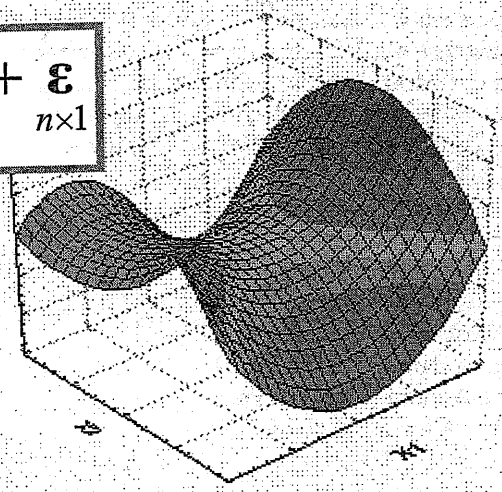
Cátedra de Métodos Cuantitativos Aplicados

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(k+1)}{\mathbf{X}} \cdot \underset{(k+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$



Modelos Estadísticos

Para las carreras de Agronomía, Ciencias Ambientales y Gestión de Agroalimentos

Facultad de Agronomía - UBA

Guía de Clases Teóricas 2008

Cartelera del curso en internet: www.agro.uba.ar/users/secmca

Cátedra de Métodos Cuantitativos Aplicados

Equipo docente de la cátedra

Ing. Agr. Susana B. Perelman (Profesora Asociada- Coordinadora de la cátedra)
Ing. Agr. Norberto J. Bartoloni (Profesor Asociado)
Lic. Mirta González (Profesora Asociada)
Ing. Agr. M. Virginia López (Profesora Adjunta)
Ing. Agr. William B. Batista (Profesor Adjunto)
Lic. Olga S. Filippini (Profesora Adjunta)
Lic. Víctor Brescia (Profesor Adjunto)
Lic. María C. Fabrizio (Profesora Adjunta)
Ing. Agr. Teresa Boca (Jefe de TP)
Ing. Diana Giorgini (Jefe de TP)
Ing. Agr. Gustavo A. Sznaider (Jefe de TP)
Dr. Karina Hodara (Jefe de TP)
Dr. Roxana Aragón (Jefe de TP)
Dr. Pablo A. Cipriotti (Jefe de TP)
Ing. Agr. Laura Puhl (Jefe de TP)
Ing. Agr. Luis Frank (Jefe de TP)
Ing. Agr. Lucas Garibaldi (Ayud. 1era)
Ing. Agr. Gonzalo Grigera (Ayud. 1era)
Lic. Fernando Biganzoli (Ayud. 1era)
Ing. Agr. Pedro Tognetti (Ayud. 1era)
Lic. Alexis Cerezo (Ayud. 1era)
Lic. Marcos Texeira (Ayud 1era)
Ing. Agr. Cecilia Conde (Ayud. 1era)
Sebastián Giedzinski (Ayudante alumno)
Erika Hirschowitz (Ayudante alumna)
Ileana Boyle (Ayudante Alumna)
Facundo Gallo (Ayudante alumno)

Pabellón Wernicke (Ver plano en http://www.agro.uba.ar/fauba/ubicacion/plano.htm)
Teléfono: 4524-8077 **Secretaria:** Ana Inés Garaño e-mail: garano@agro.uba.ar

Carreras de Agronomía, Ciencias Ambientales y Gestión de Agroalimentos.

Estadística general (cuatrimestral, 5 hs. semanales, se dicta en ambos cuatrimestres)
Modelos Estadísticos (bimestral, 6hs. semanales, se dicta en 1er y 3er bimestres)

Licenciatura en Economía y Administración Agrarias

Estadística I (cuatrimestral, 6horas semanales, 2° cuatrimestre)
Econometría (cuatrimestral, 4horas semanales, 1er cuatrimestre)
Estadística para Administradores (cuatrimestral, 3horas semanales, 1er cuatrimestre)

Licenciatura en Ciencias Ambientales

Introducción a la programación para CsAmb (bimestral, 5 horas semanales, 1er bimestre)

Carrera técnica en Floricultura

Estadística (bimestral, 2 horas semanales, 2° bimestre)

Cursos de intensificación y especialización

Estadística aplicada a la protección vegetal
Colección e Interpretación de datos

Cursos de Maestría y Doctorado en la EPG- FAUBA

Estadística Aplicada - Maestría en Agronegocios
Estadística Aplicada a la Investigación Biológica - Maestría en Biometría y Mejoramiento
Análisis Multivariado Descriptivo en Investigaciones Biológicas y Sociales- Maestría en Biometría y Mejoramiento
Teoría Estadística - Maestría en Biometría y Mejoramiento
Análisis de la Heterogeneidad de la Vegetación - Maestría en Recursos Naturales
Análisis de Datos Categóricos- Maestría en Biometría y Mejoramiento
Ecología de Poblaciones Vegetales- Maestría en Recursos Naturales

Modelos Estadísticos

Cátedra de Métodos Cuantitativos Aplicados

Duración de la materia: 8 semanas (Bimestral)
Carga horaria: 6 horas por semana: Teórico (3 horas) y Práctico (3 horas).
Asistencia: 75 % de asistencia a las clases para conservar la regularidad.

La materia tiene un **régimen de aprobación promocional**.
La evaluación es por medio de 2 (dos) exámenes parciales, 1 (un) trabajo domiciliario y trabajos adicionales que incluyen la entrega de ejercicios resueltos en cada clase práctica y resolución de ejercicios en la clase de TP.
La calificación máxima del primer parcial es de **20** (veinte) puntos y la del segundo examen integrador es de **60** (sesenta) puntos.
El Trabajo domiciliario tiene como calificación máxima **10** (diez) puntos y carácter de obligatorio para promocionar la materia.
Los trabajos adicionales (entrega de ejercicios resueltos y/o problemas realizados en clase) califican con un total máximo de **10** (diez) puntos.
- Los alumnos que en su evaluación general acumulen **70** (setenta) o más puntos, aprueben el trabajo domiciliario y cumplan con la asistencia obligatoria, promocionarán la materia,
- aquellos que acumulen entre **40** (cuarenta) y 69 (sesenta y nueve) puntos y cumplan con la asistencia obligatoria alcanzarán la regularidad y
- quienes acumulen menos de 40 puntos y/o no cumplan con la asistencia, quedarán en condición de libres.

Modalidad de los parciales

Los parciales serán a libro cerrado.
Los estudiantes pueden tener como ayuda-memoria durante el examen un resumen manuscrito confeccionado por ellos mismos (no se aceptarán fotocopias), que debe ser entregado junto con el parcial (se admite un resumen de 1 carilla para el primer parcial y de 2 para el segundo).

IMPORTANTE

En este curso, cualquier forma de engaño o plagio, ya sea que esta tenga lugar durante los exámenes o en la preparación de trabajos domiciliarios, será considerada una **falta de ética grave** y resultará en una penalización severa para todos los estudiantes involucrados. La presentación de informes idénticos (o excesivamente similares) y la utilización del trabajo de otros en los textos sin citarlos se considera plagio y, cuando sea detectado en trabajos domiciliarios obligatorios para alcanzar la promoción, los estudiantes involucrados perderán la posibilidad de promocionar la materia. La copia durante un examen implicará **calificación cero** en esa evaluación para todos los involucrados. Las actitudes solidarias y la colaboración entre estudiantes, que son fuertemente promovidas por la cátedra, excluyen explícitamente facilitar a otros el acceso al propio trabajo durante las instancias de evaluación.

Trabajos prácticos domiciliarios

Requisitos mínimos para que el trabajo sea aceptado:

1. El TP deberá contener los datos de los estudiantes y del curso al que asisten (apellidos, nombres, turno al que asisten y nombre del docente).

2. El TP será entregado en el lugar, horario y fecha límite establecida por el docente, sin excepciones. La entrega se hará en hojas impresas (no se aceptarán disquetes ni envío de archivos electrónicos).

3. El TP debe contener:

   - **Enunciado del problema.** Identificación de la problemática en cuestión, la/ s variable/ s y las hipótesis de interés.

   - **Resultados:** Tablas y gráficos brevemente comentados.

   - **Conclusión:** Se presentarán las interpretaciones de los resultados en términos del problema.

CRONOGRAMA PRIMER BIMESTRE 2008

Semana 1 (3 al 7 de Marzo)
Teórico 1: Introducción a los modelos estadísticos y el diseño de experimentos.
Práctico 1: Visita al campo experimental. Diseño de experimentos. Tratamientos, variable respuesta, unidades experimentales, repeticiones, control de la heterogeneidad en el material experimental.

Semana 2 (10 al 14 de Marzo)
Teórico 2: El análisis de varianza. Modelo lineal aditivo. Las pruebas de hipótesis. Comparaciones múltiples.
Práctico 2: Análisis de varianza de una vía. El modelo lineal y sus parámetros. Pruebas de hipótesis. Comparaciones múltiples

Semana 3 (17 al 21 de Marzo) SEMANA SANTA
Teórico y Práctico 3: El control de la heterogeneidad. Modelos asociados a los diseños completamente aleatorizado y en bloques completos aleatorizados. Análisis de varianza y pruebas de hipótesis.

Semana 4 (24 al 28 de Marzo)
Teórico 4: Violación de los supuestos del análisis de varianza: diagnóstico y alternativas para el análisis de los datos. Número de repeticiones. Integración y repaso DCA y DBCA.
Práctico 4: **Primer parcial (1/2 clase) y TP en el centro de cómputos (1/2 clase).** Entrega de consigna para el TP domiciliario.

Semana 5 (31 de Marzo al 4 de Abril)
Teórico y Práctico 5: Experimentos factoriales. Partición de la variabilidad atribuida a los tratamientos. Interacción entre factores. Modelo, supuestos y análisis. Interpretación de los resultados. Introducción a los Modelos Mixtos.

Semana 6 (7 al 11 de Abril) Entrega de TP domiciliario antes de la clase.
Teórico y Práctico 6: Análisis de Regresión Múltiple. Interpretación del Modelo. Estimación de los parámetros. Análisis de Varianza de la regresión. Pruebas de Hipótesis. Multicolinealidad.

Semana 7 (14 al 18 de Abril)
Teórico 7: Análisis de Regresión Múltiple. Evaluación del Modelo. Regresión polinómica. Regresión con variables categóricas. Introducción al análisis de Covarianza.
Práctico 7: TP en el centro de cómputos. Análisis de Regresión Múltiple.

Semana 8 (21 al 25 de Abril)
Teórico 8: Integración y Repaso.
Práctico 8: Examen integrador.

INDICE

Capítulo 1: Experimentación y Modelos Estadísticos.....Pág. 1

Introducción a la estadística experimental..... Pág. 1
Incertidumbre..... Pág. 1
Preguntas, hipótesis, predicciones y objetivos..... Pág. 2
Diseño de experimentos..... Pág. 4
Tratamientos, unidades experimentales y..... Pág. 4
variable respuesta
Repeticiones y error experimental..... Pág. 7
Independencia entre unidades experimentales..... Pág. 9
Aleatorización..... Pág. 10
Potencia..... Pág. 11
Variabilidad: su control..... Pág. 12
Modelos estadísticos..... Pág. 14
Estimación..... Pág. 17
Las pruebas de hipótesis estadísticas..... Pág. 17
Experimentos manipulativos y mensurativos:..... Pág. 18
alcances de las conclusiones

Capítulo 2: Análisis de la Varianza.....Pág. 21

Partición de la Variabilidad..... Pág. 23
Partición de los Grados de Libertad..... Pág. 25
Cuadrados medios..... Pág. 25
La distribución F de Snedecor..... Pág. 26
El Cociente F y La prueba de Hipótesis..... Pág. 26

Capítulo 3: El Modelo Lineal Aditivo.....Pág. 33

Estimación de los Parámetros del Modelo..... Pág. 34
Las hipótesis y El Modelo..... Pág. 35
Esperanzas de los Cuadrados Medios..... Pág. 38
Comparaciones Múltiples..... Pág. 39
El número de Repeticiones..... Pág. 41
Intervalos de confianza en Comparaciones Múltiples..... Pág. 41
Tasa de Descubrimientos Falsos..... Pág. 42

Capítulo 4: El Control de la Heterogeneidad.....Pág. 45

Diseño Completamente Aleatorio (DCA)..... Pág. 45
Diseño en Bloques Completos Aleatorizados (DBCA)..... Pág. 47
Modelo para el diseño en Bloques completos Aleatorios.... Pág. 48
Supuestos del Modelo..... Pág. 48
Partición de la Variabilidad total en el DBCA..... Pág. 50
Efectos Confundidos..... Pág. 52
El Control de la Heterogeneidad..... Pág. 52

Capítulo 5: Validación de los supuestos.....Pág. 55

Validación de los supuestos del ANOVA..... Pág. 55
Supuestos de homogeneidad de Varianzas y Normalidad Pág. 57
Supuesto de independencia entre bloques y tratamientos Pág. 61

Capítulo 6: Experimentos Factoriales.....Pág. 63
Experimentos Factoriales..... Pág. 63
Modelo lineal para el experimento factorial..... Pág. 67
Estimadores Puntuales de los Parámetros..... Pág. 68
Estimación de los efectos Principales y Simples..... Pág. 70
Análisis de la Varianza para un experimento factorial..... Pág. 71
Partición de los Grados de Libertad..... Pág. 72
Pruebas de Hipótesis..... Pág. 73
Prueba para el efecto de interacción..... Pág. 73
Experimentos factoriales y ciencias agropecuarias..... Pág. 77
Diseño de parcelas divididas..... Pág. 81
Factores Aleatorios y Modelos Mixtos..... Pág. 86

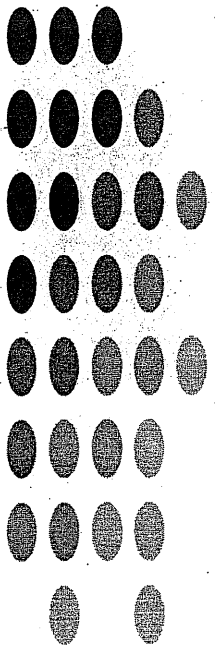
Capítulo 7: Análisis de Regresión.....Pág. 89
Análisis de regresión..... Pág. 89
Regresión lineal con una variable independiente..... Pág. 89
Modelo de regresión..... Pág. 90
Estimación de la función de regresión..... Pág. 91
Inferencia acerca de la pendiente β_1..... Pág. 92
ANOVA para el análisis de regresión..... Pág. 93
Estimación del valor esperado de Y para un valor dado de X Pág. 95
Predicción de un valor particular de Y para un valor dado de X Pág. 95
Coeficiente de determinación R^2..... Pág. 96
Regresión Múltiple..... Pág. 97
Modelo de primer orden con dos variables independientes... Pág. 97
Significado de los Coeficientes de Regresión..... Pág. 98
Modelo general de Regresión Lineal Múltiple. ..... Pág. 99
Efectos de interacción..... Pág. 99
Modelo de Regresión Lineal Múltiple en notación matricial Pág. 100
ANOVA para el análisis de regresión múltiple..... Pág. 101
Coeficiente de determinación múltiple..... Pág. 101
El problema de la multicolinealidad..... Pág. 104
Regresión polinómica..... Pág. 106
Regresión con variables explicativas categóricas..... Pág. 107
Análisis de Covarianza y Regresión con variables categóricas Pág. 110
Las hipótesis en ANCOVA y en ARVC..... Pág. 112
Independencia de la covariable respecto a los..... Pág.116
tratamientos
Los residuos como elemento de diagnóstico..... Pág. 117

Capítulo 1
Experimentación y modelos estadísticos

Las motivaciones para realizar experimentos

Los aspectos a tener en cuenta al planificar los experimentos

El modelo como síntesis de las preguntas del investigador



Introducción a la estadística experimental

Los profesionales de las ciencias agropecuarias y ambientales enfrentan en su vida cotidiana múltiples situaciones en las que deben tomar decisiones, la mayoría de las veces basándose en información incompleta. Aquellos que se dedican a la investigación y participan en la generación de nuevos conocimientos, necesitan reglas para diseñar correctamente sus estudios. Otros, que no trabajan en investigación, se ven bombardeados frecuentemente por cúmulos de nueva información que deben seleccionar y evaluar objetivamente antes de incorporarla a su práctica profesional. Unos y otros requieren las herramientas que provee la estadística para establecer el grado de fiabilidad de los resultados generados en trabajos experimentales.

En la primera mitad del siglo XX los trabajos de R. A. Fisher (Figura 1.1) sentaron las bases del diseño de los experimentos agrícolas, proyectados específicamente para determinar los efectos de uno o varios factores sobre alguna variable de interés. Los mismos fundamentos se aplican actualmente para diseñar estudios que permiten determinar, por ejemplo, el efecto de la fecha de siembra y la aplicación de fertilizante nitrogenado sobre el rendimiento de un cultivo.

Figura 1.1. Sir Ronald Aylmer Fisher (1890 Londres- 1962 Adelaida). Hizo enormes aportes a la Estadística Experimental y a la Genética Evolutiva.

En la planificación de los experimentos, a través de una metodología para la colección de los datos y su análisis, se apunta a desentrañar el complejo de variables que intervienen en el fenómeno bajo estudio, apartando a todas aquellas que ocultan la acción de los factores que se desea estudiar. Esto permite poner a prueba las hipótesis planteadas por el investigador y establecer restricciones a la interpretación arbitraria de la información obtenida.

1. Incertidumbre

Habitualmente en la vida profesional se deben tomar decisiones bajo incertidumbre. Por ejemplo, un evaluador ambiental debe decidir si la instalación de un gasoducto afecta significativamente la diversidad de especies de la selva que atraviesa, con la limitación que impone la imposibilidad de inventariar la diversidad de especies en TODA la superficie que puede ser afectada por el gasoducto y además sospechando fuertemente que los sitios no inventariados no tienen exactamente la MISMA diversidad de especies que aquellos evaluados. La estadística provee herramientas para cuantificar y disminuir la incertidumbre debida a estos dos aspectos fundamentales de la realidad: la variabilidad en las características estudiadas y la información incompleta.

Variabilidad en las características estudiadas. Uno de los rasgos más comunes de los fenómenos biológicos consiste en que sus manifestaciones no se repiten exactamente, de modo que cuando vamos a estudiarlas las mediciones que obtenemos nunca son iguales. Así, cinco plantas provenientes de semillas de girasol aparentemente idénticas,

cultivadas en macetas con el mismo suelo y ambiente, muy difícilmente acumularán la misma biomasa o presentarán la misma altura a los 20 días de la emergencia.

Esta variación observada en los fenómenos biológicos se debe a múltiples causas. Por ejemplo, la biomasa de las cinco plantas de girasol podría diferir porque, entre otras razones, las 5 plantas tienen distinto genotipo, hay pequeñas diferencias en fertilidad o humedad del suelo entre las macetas, o la ubicación en el espacio no es la misma entre las 5 macetas. Generalmente los profesionales de las ciencias agropecuarias y ambientales están interesados en estudiar la variación en el fenómeno estudiado en función de alguna/s causa/s. Sin embargo, además de las causas que se estudian, los fenómenos biológicos están afectados por otros muchos factores que es necesario considerar a la hora de interpretar los resultados obtenidos. Por ejemplo, supongamos el caso en que se desee estudiar el efecto del tipo de labranza (labranza convencional vs. siembra directa) sobre la invasión de malezas perennes en la Pampa Ondulada. La invasión de malezas podría verse afectada, no sólo por el tipo de labranza, sino también por los cultivos antecesores, por los cultivos actuales, por el tipo de suelos, por la distancia a fuentes de semillas, por la presencia de aves que las dispersen, o por el manejo diferencial de los cultivos que conducen los productores, entre otros muchos factores. En los siguientes capítulos se presentarán las herramientas que provee la estadística para disminuir la incertidumbre a partir de la cuantificación de la variabilidad.

Información incompleta. Usualmente no es factible observar todas las posibles manifestaciones del fenómeno estudiado, ya que no se pueden medir las variables de interés en todas las unidades de la población que se está investigando. Tal es el caso del evaluador ambiental que debe estudiar el impacto del gasoducto sobre la diversidad de la selva. Necesariamente debemos seleccionar un subconjunto de unidades muestrales sobre las cuales tomar los datos o realizar las observaciones y a partir de esa información incompleta extraer conclusiones acerca de la población.

2. Preguntas, hipótesis, predicciones y objetivos

La conducción de experimentos es una manera adecuada de responder preguntas acerca de la naturaleza, como las problemáticas planteadas en los ejemplos anteriores acerca de la sustentabilidad de los cultivos, de la diversidad de especies en la selva o de la invasión de malezas perennes en los cultivos. En muchos casos, los profesionales tienen posibles respuestas a estas preguntas antes de realizar el ensayo (i.e. **hipótesis**). Las hipótesis son afirmaciones sobre la realidad que solo pueden verificarse de manera indirecta, es decir, por alguna de sus consecuencias (i.e. **predicciones**). Entre las diversas formas que existen para generar conocimiento, el uso del método hipotético-deductivo ofrece la ventaja de permitir poner a prueba una idea que fue elaborada antes de realizar el estudio (hipótesis, Farji-Brener 2003). Este método consiste en deducir consecuencias concretas (i.e. generear predicciones derivadas de la hipótesis) sobre un determinado objeto de estudio bajo el supuesto de que nuestra afirmación o idea (hipótesis) es verdadera. De esta forma, al comparar los resultados obtenidos en nuestro ensayo contra los resultados que esperaríamos obtener si nuestra hipótesis fuese correcta, se puede poner a prueba la veracidad de la hipótesis y así avanzar en el conocimiento del problema planteado (Figura 1.2).

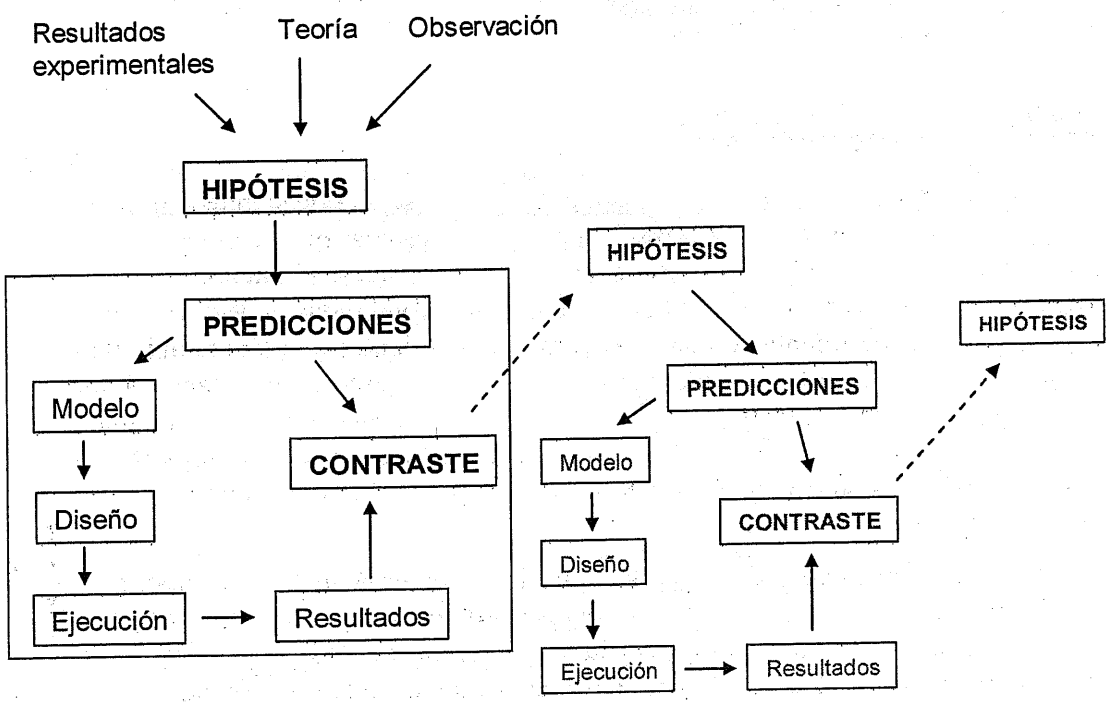


Figura 1.2. Esquema general de los componentes de un programa de investigación. El recuadro marca los límites del experimento. Del contraste de las predicciones de las hipótesis surgen conclusiones que generan nuevas hipótesis.

Por ejemplo, si se postula que la siembra directa en la Pampa Ondulada favorece la invasión de los campos con malezas perennes, la predicción derivada de esta hipótesis sería que en lotes con larga historia de siembra directa se encontrará una mayor abundancia de malezas perennes que en aquellos lotes que durante el mismo tiempo se cultivaron con labranza convencional. Un experimento correctamente diseñado permitiría verificar si esta predicción se sostiene en los lotes cultivados de la Pampa Ondulada. Para ello se deberán contestar varias preguntas de este tipo:

- ¿Cómo se va a medir el efecto de la invasión con malezas perennes?
- Además del tipo de labranza (siembra directa vs. labranza convencional), ¿qué otros factores podrían afectar las respuestas que se van a analizar?
- ¿Cuáles de esos factores se van a considerar en el estudio?
- ¿Cuál es la población a la que se van a referir las conclusiones?
- ¿Cuál sería la forma de analizar los resultados?
- ¿Qué diferencias en la abundancia de malezas se van a considerar importantes?
- ¿Cuántas veces se deberían repetir las observaciones para obtener resultados confiables?

El **objetivo** enuncia qué es lo que se pondrá a prueba en la investigación y cuál es el objeto de estudio. En el ejemplo de la invasión de malezas, el objetivo del trabajo

podría ser: Estudiar el efecto del tipo de labranza sobre la invasión de malezas perennes en lotes cultivados de la Pampa Ondulada.

3. Diseño de experimentos

Realizar un buen diseño es planificar el experimento de manera que se pueda obtener la información adecuada al menor costo posible. Es decir, asegurar que se tomarán los datos que conduzcan a decisiones válidas en relación al problema planteado, minimizando los costos en tiempo y esfuerzo. Las reglas contenidas en los principios básicos del diseño experimental ayudan a generar conocimientos de manera objetiva, precisa, sistemática y repetible. La planificación del experimento es una etapa crítica e irreversible en este proceso, ya que ningún análisis posterior puede convertir un mal diseño en uno bueno. Si el experimento no es correctamente diseñado, no importa cuan bueno sea el análisis posterior de los datos, ya que no se podrá cumplir con el objetivo del estudio (Eberhardt & Thomas 1991).

Los experimentos conducidos en laboratorio, en invernáculos o a campo, no difieren en su estructura lógica; así sea que se trate de estudiar el efecto de la calidad de la luz sobre la germinación de semillas de chamico en cajas de Petri, el crecimiento de plantas de girasol en macetas con distinta fertilidad de suelos o la relación entre el diámetro de los troncos y el volumen de madera en una plantación de pinos (Figura 1.3). A pesar de las diferencias de contexto, todos estos estudios encuentran dificultades semejantes para aislar los efectos buscados, porque la interpretación de cualquier fenómeno biológico requiere prestar atención a los obstáculos antes mencionados: la variabilidad en las respuestas, las causas múltiples que las afectan y la incapacidad de observar la población completa. En la consideración de estos aspectos se fundamentan los principios básicos del diseño de los experimentos: la *repetición*, la *aleatorización* y el *control de la heterogeneidad*, que serán discutidos a continuación (Figura 1.4, secciones 5 a 9). Previamente presentaremos la definición de *tratamientos, unidades experimentales* y *variable respuesta*, elementos indispensables para esa discusión.

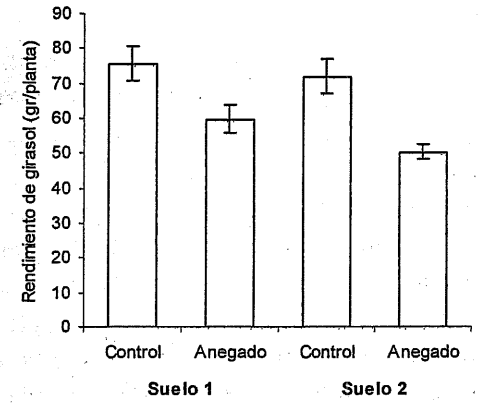
4. Tratamientos, unidades experimentales y variable respuesta.

Se denomina **tratamientos** a los diferentes procesos cuyos efectos van a ser medidos y comparados. Por ejemplo, podríamos estar interesados en comparar el rendimiento de cuatro híbridos de maíz, o el efecto de tres niveles de riego (sequía, inundación y riego normal) sobre el crecimiento de un cultivo, o el efecto de tres herbicidas diferentes para controlar malezas dicotiledóneas, o la efectividad de dos vacunas distintas para prevenir la aftosa. En todos estos casos el experimento consiste en aplicar a cada *unidad experimental* (parcela, maceta, animal, establecimiento, planta, etc) uno de los distintos tratamientos a comparar y medir la *respuesta* que producen, para poder luego estimar sus efectos.

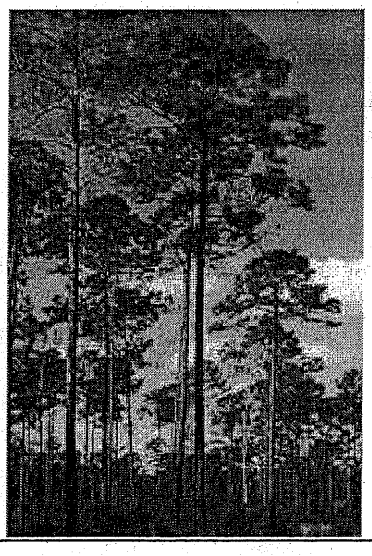
Una **unidad experimental** es la mínima porción de material experimental que puede recibir un tratamiento (y que, en consecuencia, podría haber recibido cualquier otro) y sobre la cual se observará una respuesta para evaluar el efecto del tratamiento. En los ejemplos de la Figura 1.3 las unidades experimentales son respectivamente: cada maceta con una planta de girasol, cada árbol de la plantación de pinos y cada caja de Petri con 20 semillas de chamico. Las unidades experimentales pueden tener distinto

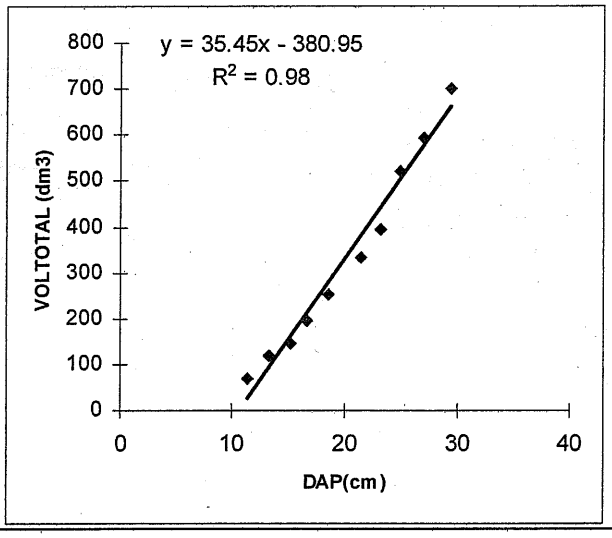
Figura 1.3. Resultados de experimentos conducidos en macetas (girasol), plantación (pino) y laboratorio (chamico).

Efecto del anegamiento sobre el rendimiento de girasol en distintos suelos.

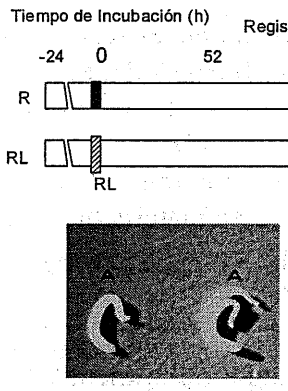


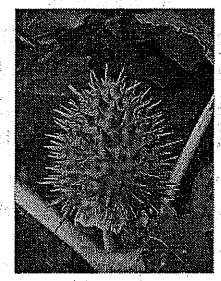
Relación entre el volumen individual (dm^3/árbol) y el diámetro a la altura del pecho (DAP en cm) en árboles de *Pinus elliottii* en la localidad de Esquina (Corrientes).

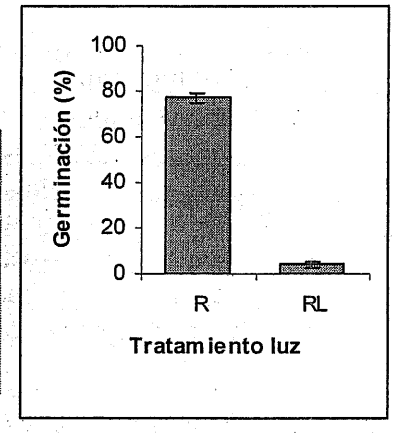




Germinación de semillas de *Datura ferox* (chamico) en respuesta a la calidad de la luz.







tamaño y forma, dependiendo por ejemplo de la especie en estudio, de la heterogeneidad ambiental presente o del tipo de tratamiento a asignar.

Las preguntas que intentamos responder mediante un experimento o un plan de investigación se refieren generalmente a una población, que debe ser explícitamente identificada al momento de planificar el estudio. Habitualmente es improbable que podamos observar o estudiar todas las unidades experimentales de la población, por lo tanto, se obtiene una **muestra** representativa de la población. Es decir, se selecciona un *subconjunto* de unidades experimentales (muestra) del *conjunto* de todas las unidades experimentales (población). Por ejemplo, si estamos interesados en estudiar el diámetro de los troncos de los árboles de álamo cultivados en el delta del Río Paraná (población de referencia), la unidad experimental será cada árbol. Una muestra podría ser un conjunto de 150 árboles de álamo seleccionados al azar en las explotaciones comerciales del delta del Paraná. Como se mencionó anteriormente, habitualmente nos manejamos con información incompleta ya que no contamos con mediciones realizadas en todas las unidades que constituyen la población. Es decir que seleccionamos una muestra para tomar decisiones o realizar conjeturas, métodos estadísticos mediante, respecto de la población completa.

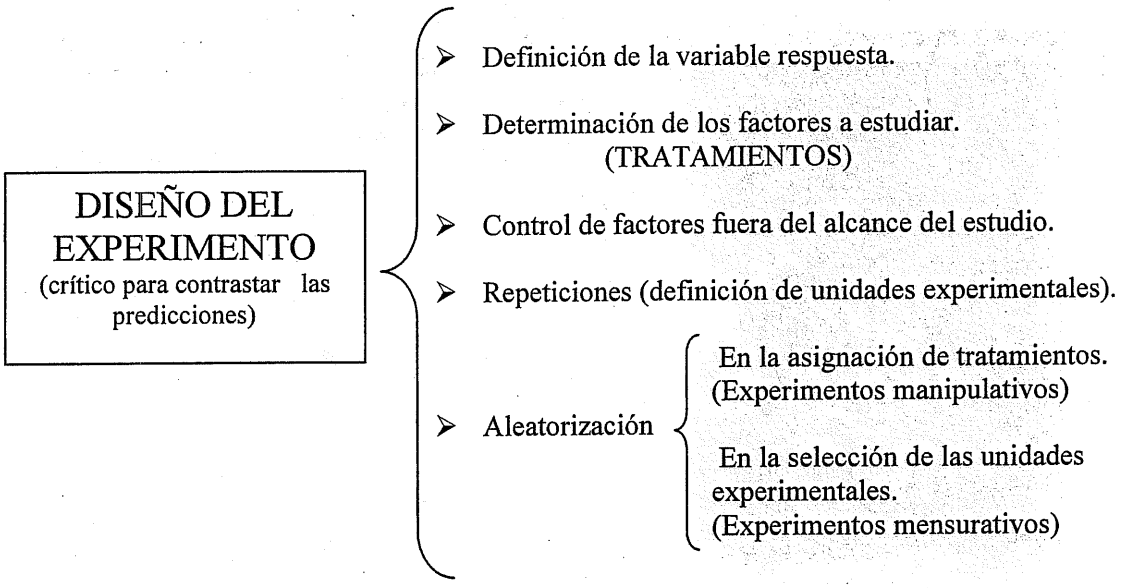

DISEÑO DEL EXPERIMENTO
(crítico para contrastar las predicciones)
Definición de la variable respuesta.
Determinación de los factores a estudiar. (TRATAMIENTOS)
Control de factores fuera del alcance del estudio.
Repeticiones (definición de unidades experimentales).
Aleatorización
En la asignación de tratamientos. (Experimentos manipulativos)
En la selección de las unidades experimentales. (Experimentos mensurativos)

Figura 1.4. Principales elementos a considerar en el diseño experimental.

Finalmente, es muy importante que las unidades experimentales sean representativas de la población sobre la que se han fijado los objetivos del estudio. Por ejemplo, los pastizales inundables del SE de la provincia de Buenos Aires se encuentran bajo el área de influencia de la estación experimental del INTA en Balcarce. Sin embargo, si se realizaran experimentos de manejo de pasturas en el mismo sitio de la estación experimental, que tiene suelos profundos y muy ricos en materia orgánica, no correspondería extrapolar los resultados para asesorar a los productores del partido de Castelli, quienes manejan pasturas en ambientes con suelos someros hidro o halomórficos (con problemas de inundación y salinización).

Cualquier pregunta que queramos responder mediante un experimento implica realizar alguna observación o medición (**variable respuesta**) sobre cada una de las unidades experimentales de la muestra. De este modo, en el ejemplo de los álamos,

podríamos estar interesados en observar la altura, el diámetro a la altura del pecho, el volumen del árbol o la rectitud del fuste entre otros, dependiendo del objetivo del estudio. La variable observada o de respuesta puede ser **cuantitativa** o **cualitativa**. Una variable cuantitativa asume valores numéricos que poseen un orden inherente, tal como los diámetros de tallos de plantas o las edades de un grupo de empleados. A su vez, una variable cuantitativa puede ser **cuantitativa discreta** (número de hojas por planta, cantidad de pétalos por flor) o **cuantitativa continua** (diámetros de tallos, pesos de semillas). Una variable cualitativa o categórica, es aquella cuyos valores corresponden a una u otra categoría de un conjunto de categorías que son mutuamente excluyentes. Por ejemplo, infectada o no infectada son dos posibles categorías de la variable plantas examinadas por su sanidad. Igualmente, rojo, azul y blanco podrían ser tres posibles categorías para la variable color de los pétalos.

Ocasionalmente los tratamientos no pueden ser aplicados o manipulados por el investigador, sino que ya vienen incluidos en las unidades experimentales (Hulbert 1984). Por ejemplo, si se desea comparar la facilidad de incorporación de nuevas prácticas agropecuarias entre productores con distinto nivel de educación es obvio que los tratamientos (educación primaria, secundaria y universitaria, por ejemplo) no pueden ser aplicados a los productores por el investigador. De igual forma, si se quiere estudiar el crecimiento de árboles de lenga en relación con la exposición (norte o sur) de las laderas en las que se encuentran los bosques, el factor exposición no puede ser aplicado, sino que las unidades experimentales ya están asignadas a un determinado tratamiento por su ubicación espacial. En estos casos, los tratamientos no se manipulan, sino que simplemente se registran sus valores; se aprovecha la variabilidad natural en las características de interés para ver si se ajustan o no a lo que predicen las hipótesis que se formularon. Algunos autores denominan a este tipo de tratamientos **factores de clasificación** y a los experimentos que los estudian **mensurativos** en contraposición a los **experimentos manipulativos**, en los cuales el investigador aplica los tratamientos. La distinción no tiene relación con la precisión de los instrumentos de medida utilizados en uno y otro caso, ni con el lugar de la toma de datos (campo o laboratorio), sino sólo con el hecho de que se pueda aplicar o no un tratamiento a cualquier unidad experimental (Figura 1.4).

5. Repeticiones y error experimental

La acción de repetir consiste en asignar cada uno de los tratamientos a varias unidades experimentales independientes. Dado que ningún animal, planta o parcela es idéntico a otro, siempre habrá alguna variación en las respuestas que no ha sido controlada, aún cuando el experimento sea conducido muy cuidadosamente. En consecuencia, los resultados de los experimentos están afectados no sólo por la acción de los tratamientos que aplicamos, sino también por las causas múltiples no controladas que tienden a encubrir sus efectos (Figura 1.5). Estas diferencias en la respuesta entre unidades que recibieron el mismo tratamiento, que en consecuencia son diferencias no debidas al efecto de los tratamientos, se denominan errores experimentales. En otras palabras, el error experimental es una medida del fracaso en alcanzar resultados idénticos en unidades experimentales que fueron tratadas idénticamente.

La repetición de unidades experimentales independientes expuestas al mismo tratamiento permite estimar el error experimental y utilizarlo como unidad básica de medida para evaluar el significado de las diferencias observadas entre tratamientos. "Error" no significa que nos hayamos confundido, el error experimental representa el

efecto del resto de los factores (que no son de interés para el estudio) sobre la variable respuesta.

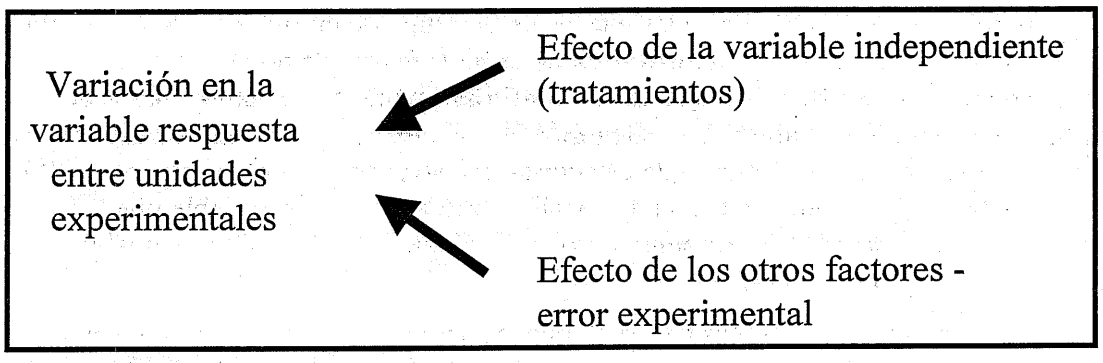

Variación en la variable respuesta entre unidades experimentales
Efecto de la variable independiente (tratamientos)
Efecto de los otros factores - error experimental

Figura 1.5. Determinantes de la variación entre distintas unidades en la variable respuesta.

Por ejemplo, si se desea comparar el rendimiento de dos híbridos de maíz (híbridos A y B), 5 repeticiones por tratamiento implicaría asignar 5 parcelas a cada híbrido, de modo que el experimento contaría con 10 unidades experimentales en total (Figura 1.6). Como podemos observar, existe variación en el rendimiento entre las unidades experimentales, es decir, las 10 unidades experimentales no presentan el mismo rendimiento. Una parte de esa variación se debe al efecto de los tratamientos, o sea, a las diferencias entre los dos híbridos. La otra parte de esa variación se debe al error experimental, es decir, a diferencias entre las parcelas en cuanto a todos los otros factores que afectan al rendimiento (topografía, historia de uso, manejo, etc.). Por lo tanto, el error experimental abarcaría las diferencias en rendimiento observadas entre las parcelas que fueron cultivadas con el mismo híbrido. De este modo, si observamos dos parcelas con el híbrido A, el mayor rendimiento observado en la parcela 3 con respecto a la 4 no se debe a diferencias entre los híbridos. En cambio, este mayor rendimiento si se debe a otros factores, por ejemplo, tal vez la parcela 3 tuvo menor historia de uso o mayor humedad que la parcela 4.

Parcela nº	Híbrido	Rendimiento
1	A	88
2	A	60
3	A	89
4	A	79
5	A	68
6	B	50
7	B	56
8	B	65
9	B	77
10	B	51

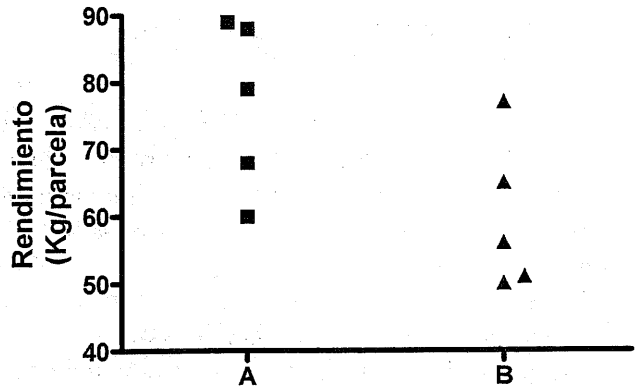


Figura 1.6. El rendimiento observado en cada parcela varía según el híbrido sembrado (tratamientos) y según otros factores (error experimental). A= Híbrido A, B= Híbrido B.

Cuando los tratamientos no se encuentran repetidos en varias unidades experimentales, es imposible discernir si las diferencias observadas se deben a

diferencias entre tratamientos o a diferencias intrínsecas entre las unidades experimentales. Esto nos impide evaluar de manera concluyente el efecto de los tratamientos. Por ejemplo, supongamos el caso en que no existen diferencias entre esos dos híbridos, y realizamos un ensayo con 5 repeticiones. Los resultados podrían tener un aspecto similar a los de la Figura 1.7a. En el caso en que hubiéramos realizado este experimento con una sola repetición, es decir una parcela con el hibrido A y otra parcela con el híbrido B, podríamos haber observado algo similar a la Figura 1.7b ya que nunca dos unidades experimentales son iguales. La diferencia en el rendimiento entre esas dos parcelas, puede deberse a una diferencia desconocida de fertilidad de suelo entre los lotes o a una genuina diferencia de potencial de rendimiento entre los híbridos. El agrónomo realizando el ensayo interesado en detectar diferencias entre los híbridos podría sugerir incorrectamente que el híbrido B tiene un mayor rendimiento que el híbrido A. Obviamente, el agrónomo que conduce un estudio no está interesado en las diferencias de rendimiento entre los dos híbridos en ***esos*** dos lotes en particular: él desea que sus resultados puedan ser aplicados a muchos otros lotes semejantes que no participan de su ensayo. En consecuencia, necesita alguna manera de discernir si las diferencias observadas se encuentran dentro de los límites del error experimental o si, por el contrario, son suficientemente grandes como para no poder ser explicadas meramente por azar. O sea, cuando no existe un método para estimar el error experimental, no hay manera de determinar si las diferencias observadas indican un efecto puro de los tratamientos o si se deben a la variación inherente en el material al que fueron aplicados. En este caso, se dice que hay **efectos confundidos** ya que no se pueden separar los efectos de los tratamientos de los efectos de otros factores que no son de interés en el estudio (error experimental).

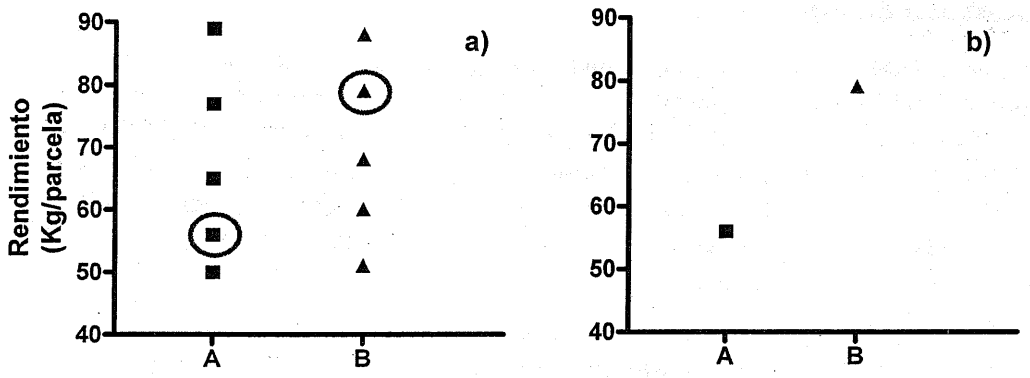


Figura 1.7. Cuando los tratamientos no son repetidos en varias unidades experimentales, no se puede discernir si las diferencias observadas se deben a diferencias entre los tratamientos o a diferencias intrínsecas entre las unidades experimentales. A=Híbrido A; B=Híbrido B.

6. Independencia entre unidades experimentales

En el contexto del diseño de experimentos se entiende por independencia que el valor observado para la variable respuesta en una unidad experimental no sea afectado o no esté relacionado con los valores de otras unidades experimentales. Dicho de otro modo, las observaciones independientes no están correlacionadas. En el diseño de un experimento nos interesa que:

- El efecto de los tratamientos sea independiente del efecto de otros factores (i.e. las unidades experimentales asignadas a un mismo tratamiento deben ser independientes entre sí).
- Las unidades experimentales asignadas a distintos tratamientos sean independientes entre sí (i.e. los tratamientos deben ser independientes entre sí).

Por ejemplo, supóngase que se desea probar el efecto de un herbicida de contacto aplicado por fumigación (dos tratamientos, con y sin herbicida) sobre el porcentaje de cobertura de una maleza en lotes sembrados con un cultivo (la unidad experimental es cada lote). En el primer caso, una falta de independencia implicaría que alguno de los tratamientos se vea favorecido o perjudicado en la asignación de unidades experimentales. Por ejemplo, que los lotes o parcelas tratados con el herbicida presenten una menor historia de uso y por lo tanto una menor cobertura de la maleza previa a la aplicación del herbicida, que aquellos lotes a los cuales no se les aplicará el herbicida. En el segundo caso, una falta de independencia podría darse, por ejemplo, si las parcelas estuvieran muy cerca unas de otras y el herbicida se aplica por fumigación en un día con alta velocidad de viento. En este caso, el herbicida llegaría a las parcelas que no deberían recibirlo debido a la deriva provocada por el viento. De este modo, las acciones tomadas en las unidades experimentales correspondientes al tratamiento con herbicida afectaron los resultados de las unidades experimentales correspondientes al tratamiento sin herbicida. Por lo tanto, no hay independencia entre las unidades experimentales asignadas a distintos tratamientos. El investigador debe realizar al menos dos tareas para favorecer la independencia entre unidades experimentales, ellas son la aleatorización y el control del experimento (i.e. que las parcelas "testigo" no reciban el herbicida debido a la deriva provocada por el viento).

7. Aleatorización

Para evitar el sesgo en las comparaciones entre medias de tratamientos, es necesario asegurar que un tratamiento particular no resulte favorecido en forma consistente en repeticiones sucesivas por alguna fuente desconocida de variación. Por lo tanto, un supuesto importante de cualquier experimento, es que las unidades experimentales no presentan diferencias entre los tratamientos anteriores a la aplicación de los mismos. Es decir, las **condiciones iniciales** son las mismas para las unidades experimentales que reciben los distintos tratamientos.

En los experimentos manipulativos, *cada tratamiento debe tener igual oportunidad de ser asignado a cualquier unidad experimental*. Al asignar al azar los tratamientos a las unidades se obtiene un procedimiento que logra, en promedio, una distribución equitativa de los efectos de múltiples factores desconocidos. De esta forma, la aleatorización, combinada con la repetición, son requisitos para obtener una medida válida del error experimental. La aleatorización permite que las decisiones acerca de la asignación de las unidades experimentales a diversos tratamientos no sean el producto de alguna distorsión, regla o parcialidad arbitraria y sistemática. La aleatorización evita que una arbitrariedad seguida sistemáticamente por el investigador o por alguien más, haga tendenciosos los resultados.

En los experimentos mensurativos, donde los tratamientos son inherentes al material, la aleatorización interviene en la selección de las unidades experimentales sobre las que se toman los datos u observaciones. El muestreo es necesario porque comúnmente no se pueden medir las variables en todas las unidades de las poblaciones

que estamos comparando. De esto se deduce de forma inmediata que un buen muestreo es aquel en el que la muestra es representativa del fenómeno o de la población estudiada, de manera tal que los resultados obtenidos se puedan extender a otras situaciones de la misma población que no van a ser medidas. El mejor método para garantizar la independencia de las unidades experimentales en lo que respecta a factores no controlados, es seleccionar al azar, de entre todas las posibles, las unidades que se van a medir.

La aleatorización es imprescindible pero no suficiente para alcanzar la independencia. Además de aleatorizar la asignación de los tratamientos a las unidades, el experimentador debe realizar todas las acciones que considere necesarias para garantizar la independencia entre las unidades experimentales. Retornando al ejemplo del herbicida de contacto, es importante que la aplicación por fumigación, así como cualquier acción desarrollada en las parcelas con el tratamiento herbicida, no afecte los resultados de las parcelas que no recibirán el herbicida (la recíproca es igualmente válida). En este caso, el experimentador debe aplicar el herbicida de manera cuidadosa, en un día con baja velocidad de viento y las parcelas deben estar lo suficientemente alejadas entre sí. En general, un buen control experimental implica una adecuada separación espacial y temporal entre las unidades experimentales.

8. Potencia

La **potencia** es la capacidad de un ensayo de detectar diferencias significativas entre tratamientos. Ensayos con gran potencia pueden detectar diferencias muy pequeñas entre tratamientos, mientras que ensayos con poca potencia solamente pueden detectar diferencias groseras entre tratamientos. La herramienta más conveniente que tiene el experimentador para dar a sus ensayos mayor o menor potencia es modificar el **número de repeticiones**. Un experimento realizado con repeticiones independientes, pero con pocas repeticiones, puede resultar inútil para solucionar un determinado problema. Por tradición, en los experimentos agrícolas se han empleado tres o cuatro repeticiones sin considerar la importante influencia que tiene esta decisión sobre la capacidad del experimento para detectar diferencias significativas entre los tratamientos. Al incrementar el número de repeticiones aumenta siempre la potencia, aunque simultáneamente aumentan el esfuerzo de la experimentación, de manera que resulta de especial importancia encontrar el número conveniente de repeticiones para cada experimento. Por ejemplo, supóngase que se desea evaluar el efecto del raleo (remoción de árboles) sobre el incremento en diámetro de árboles de lenga (*Nothofagus pumilio*) en bosques naturales. Para ello, se seleccionarán parcelas de bosque natural, a las que se les aplicará al azar una de tres prácticas de manejo: a) sin raleo, b) raleo bajo y c) raleo alto. ¿Cuántas repeticiones deben realizarse? Es decir, ¿cuántas parcelas deben ser asignadas a cada uno de los tratamientos? La respuesta a esta pregunta depende de las características de la población con la que se está trabajando así como de los intereses/objetivos del investigador. A continuación se discutirán conceptualmente algunos de estos aspectos. En capítulos posteriores se aprenderá a calcular el número de repeticiones necesario para un experimento.

En síntesis, al diseñar un experimento se debe tener en claro:

- que el experimento necesita repeticiones de las unidades experimentales.
- el tipo de experimento a realizar (manipulativo o mensurativo).
- que las unidades experimentales deben ser independientes.
- cuántas repeticiones debe tener el experimento.

9. Variabilidad: su control

Cuanto mayor sea el error experimental, menor será la potencia de un ensayo. Dicho de otro modo, a mayor error experimental, mayor será el número de repeticiones necesario para alcanzar una determinada potencia. El rubro "error experimental" abarca todas las causas de variación que no se controlaron durante el ensayo. Entre éstas algunas no se controlaron porque son incontrolables por naturaleza. Todo material biológico, por homogéneo que sea, presenta una cierta fluctuación cuyos factores determinantes no se conocen y son por lo tanto incontrolables. Además, las observaciones están sujetas a errores de medición, que deben ser minimizados dentro de lo posible. En cambio, otras fuentes de variabilidad podrían ser controladas. Por ejemplo, en experimentos con animales se conoce que ciertos caracteres cuantitativos difieren entre el macho y la hembra, o entre animales de distintas camadas o de distinta edad o de diferente raza. En un diseño completamente al azar que utiliza animales de distintas edades, camadas, razas y sexos, toda esa variación contribuye al rubro del error.

Una alternativa para reducir el error experimental es aplicar los tratamientos sobre animales *tan homogéneos como sea posible* (lo mismo se aplica a parcelas, plantas, empresas o cualquier otro tipo de unidades experimentales). En esos casos se elimina la heterogeneidad *antes*, y se aplica un diseño completamente aleatorio. Sin embargo, se debe tener en cuenta que las conclusiones de tales ensayos sólo serán aplicables a condiciones tan restringidas como las que se establecieron al seleccionar el material experimental. Es decir que *al acotar la variación del material que participa en el ensayo se restringe la población de referencia a la que apuntan las hipótesis*. Por ejemplo, se puede concluir que una ración produce el mayor aumento de peso en cerdos de determinada raza, edad, sexo, etc.

En resumen, existe un compromiso entre:

- incluir un amplio rango de condiciones, con el fin de hacer el experimento más general.
- restringir el rango de condiciones, con el fin de reducir la variabilidad y aumentar la potencia.

Uno de los descubrimientos más originales de la estadística es que se puede eliminar del error experimental un factor controlable de variación *después* de haber realizado el ensayo usando material no homogéneo. Para que esto sea posible se debe conocer *a priori* la heterogeneidad entre unidades y diseñar consecuentemente un experimento con aleatorización restringida, por ejemplo con un **diseño en bloques**.

Estos diseños permiten poner a prueba hipótesis referidas a poblaciones más heterogéneas y, en consecuencia, más amplias, sin provocar aumentos del error experimental que resten potencia al experimento. El más sencillo y más ampliamente utilizado de los diseños que controlan la heterogeneidad de esta manera se denomina **diseño en bloques completos** y se verá en detalle en los capítulos siguientes. Aquí sólo presentaremos su justificación conceptual.

Muchas veces el experimentador adquiere habilidad para predecir con cierta precisión el comportamiento de su material experimental: conoce que algunos lotes son más rendidores que otros, que algunos animales podrían aumentar de peso más rápidamente que otros, que alguna ubicación dentro del invernáculo recibe luz natural en un período más largo durante el día, etc. Todo conocimiento previo acerca de la heterogeneidad del material experimental puede servir para aumentar la potencia del experimento. En estos casos, el procedimiento a seguir comienza por clasificar a las unidades experimentales en conjuntos homogéneos y restringir la aleatorización de los tratamientos, realizándola por separado dentro de cada uno de esos grupos. Cada grupo conformado por un conjunto de unidades experimentales homogéneas se denomina **bloque**. Cuando se agrupan las unidades en bloques de manera tal que la variación entre las unidades de un bloque sea menor que la variación entre unidades de bloques diferentes, y se asignan al azar los tratamientos dentro de cada bloque (téngase en cuenta que un bloque contiene una repetición completa de todos los tratamientos), es posible comparar las respuestas a los tratamientos entre unidades del mismo bloque. Así, las variaciones de un bloque al otro no contribuyen al error experimental, en consecuencia, la potencia del experimento aumenta como resultado de reducir el error experimental.

Desde un punto de vista matemático el tratamiento que se hace de los factores-bloque es el mismo que el de los factores-tratamiento, pero su significado dentro del modelo es diferente. Hay interés en conocer la influencia de un factor-tratamiento sobre la variable respuesta mientras que no hay interés en conocer la influencia de un factor-bloque sino que éste es incorporado al diseño del experimento para disminuir el efecto residual del modelo (el error experimental) y así aumentar la potencia.

Si la causa de la heterogeneidad es una propiedad cuantitativa de cada una de las unidades experimentales que puede ser medida antes de realizar el experimento (i.e. el peso inicial de los animales, la hora de la medición cuando se evalúa el estado hídrico de las plantas, la presión sanguínea de un paciente en un experimento médico, la acidez del suelo de las parcelas en un experimento de fertilización fosforada, etc.), el factor se denomina **covariable** y permite también disminuir el error experimental y aumentar la potencia. La manera de estimar y aislar el efecto de **covariables** la estudiaremos en el capítulo 9.

En síntesis, las posibles fuentes de variación de un experimento son:

FUENTE DE VARIACIÓN	TIPO
Debida a las condiciones de interés (Tratamientos)	Planificada y sistemática
Debida al resto de condiciones controladas (Bloques y covariables)	Planificada y sistemática
Debida a condiciones no controladas (Error experimental)	No planificada y aleatoria

10. Modelos estadísticos

Una vez establecido el objetivo del estudio, se elige un **modelo** que representa el comportamiento de la variable respuesta como función de las variables explicativas.

$$\text{Respuesta observada} = \underbrace{\text{modelo matemático}}_{f(\text{variables explicativas})} + \underbrace{\text{error}}_{\text{error experimental}}$$

En este curso sólo estudiaremos modelos que tienen una sola variable respuesta y de naturaleza cuantitativa. Primero veremos aquéllos donde la porción explicativa del modelo está compuesta por variables cualitativas, que también se conocen como tratamientos o variables de clasificación porque dividen a las observaciones en clases. A continuación se presentan algunos ejemplos.

Variable explicativa (predictora o independiente) **CUALITATIVA**	Variable respuesta o dependiente **CUANTITATIVA**
Cultivo antecesor	Nivel de nutrientes en el suelo
Híbridos	Rendimiento
Tipo de labranza	Abundancia de malezas
Razas de terneros	Producción anual de carne
Tipo de ordeñe (manual o mecánico)	Incidencia de mastitis en cabras
Calidad de la luz	Germinación de semillas de chamico
Especie forrajera	Biomasa de la pastura
Nivel de educación	Gasto en alimentos

Los modelos con variables predictoras cuantitativas se estudiarán en el capítulo de *Análisis de Regresión*. A continuación se presentan algunos ejemplos.

Variable explicativa (predictora o independiente) **CUANTITATIVA**	Variable respuesta o dependiente **CUANTITATIVA**
Altura de los estudiantes	Peso de los estudiantes
Precipitación media anual	Productividad de pastizales
Cobertura del cultivo en floración	Rendimiento del cultivo
Densidad de pastos en la estepa	Producción anual de carne
Nivel de fertilidad del suelo	Rendimiento de trigo
Intensidad de luz	Tasa de fotosíntesis de una hoja
Tiempo desde emergencia	Biomasa del cultivo
Diámetro del tronco	Volumen de madera del árbol
Ingreso mensual familiar	Gasto en alimentos

El modelo matemático debe indicar la relación que se supone que existe entre la

variable respuesta y las principales fuentes de variación a estudiar. El modelo más habitual es el **modelo lineal**, como los siguientes:

RESPUESTA = Constante + Efecto de tratamientos + error

Si las unidades experimentales se agrupan en *bloques*:

RESPUESTA = Constante + Efecto de bloques + Efecto de tratamientos + error

A veces los tratamientos que se comparan en un ensayo son, a su vez, combinaciones de niveles de distintos factores. Por ejemplo si se desea estudiar el efecto de la fertilización nitrogenada sobre el nivel de proteína en grano en cultivos de trigo con y sin riego. En este caso, la fertilización es un factor que podría, por ejemplo, tener tres niveles (sin agregado de fertilizante, con dosis normal y con dosis alta) y el riego es otro factor que podría tener, por ejemplo, dos niveles (con y sin riego). Entonces las seis combinaciones de los niveles de ambos factores (ver tabla) constituyen los tratamientos y el porcentaje de proteína en grano es la variable respuesta.

FACTOR 1: Fertilizante 3 Niveles	FACTOR 2: Riego 2 Niveles	**TRATAMIENTOS** (Combinaciones de los factores)
Sin agregado	Con riego	1. sin fertilizante- con riego
	Sin riego	2. sin fertilizante- sin riego
Dosis normal	Con riego	3. fertilización normal- con riego
	Sin riego	4. fertilización normal- sin riego
Dosis alta	Con riego	5. fertilización alta- con riego
	Sin riego	6. fertilización alta- sin riego

Los estudios de este tipo, donde la respuesta de un factor se prueba a diferentes niveles de otro factor en el mismo experimento se conocen como **experimentos factoriales**, son muy utilizados en las ciencias agropecuarias y presentan interesantes propiedades que estudiaremos más adelante. Por ejemplo, si los tratamientos son *combinaciones de niveles de dos factores* A y B el modelo podría ser:

RESPUESTA = Constante + Efecto de A + Efecto de B + Interacción AxB + error

En estos modelos la respuesta viene dada por una suma de términos que representan las principales fuentes de variación planificadas más un término residual debido a las fuentes de variación aleatoria no planificada. Los modelos que se estudian en este curso se ajustan a esta forma general.

Si la variación explicada por el modelo y el residuo que queda sin explicar, constituyen fuentes independientes, se puede separar información de error. Obviamente el modelo será mejor si deja relativamente poco sin explicar. La intuición sugiere que el "mejor" modelo será el que arroje el menor promedio de errores o *residuales* (diferencia entre valores observados y valores predichos por el modelo).

En la formulación de cada modelo deben aparecer los **parámetros** poblacionales relevantes para el objetivo que impulsa el experimento, ya que el procedimiento que

permite contestar las preguntas de interés es la prueba de hipótesis referida a los parámetros del modelo. De esta forma el modelo estadístico vincula los objetivos del estudio con el diseño del experimento y el análisis de los datos (Figura 1.8.). La estimación de los parámetros que caracterizan a la población de referencia permite evaluar la evidencia que proveen los datos respecto a las predicciones del investigador (Figura 1.2). Los parámetros poblacionales son elementos fundamentales del modelo estadístico que, como se verá en detalle más adelante, sintetizan de manera formal las relaciones entre las variables bajo estudio. Es así que la selección del modelo estadístico más adecuado para cada caso estará determinada por los objetivos del estudio, y tanto el diseño como la conducción de los experimentos apuntarán a obtener la información necesaria para estimar correctamente los parámetros del modelo. Los modelos son "estadísticos" cuando tienen en cuenta la variación entre las unidades experimentales de la población y se pueden asociar probabilidades a los diferentes valores factibles de obtener. Estos conceptos serán discutidos en detalle en los próximos capítulos.

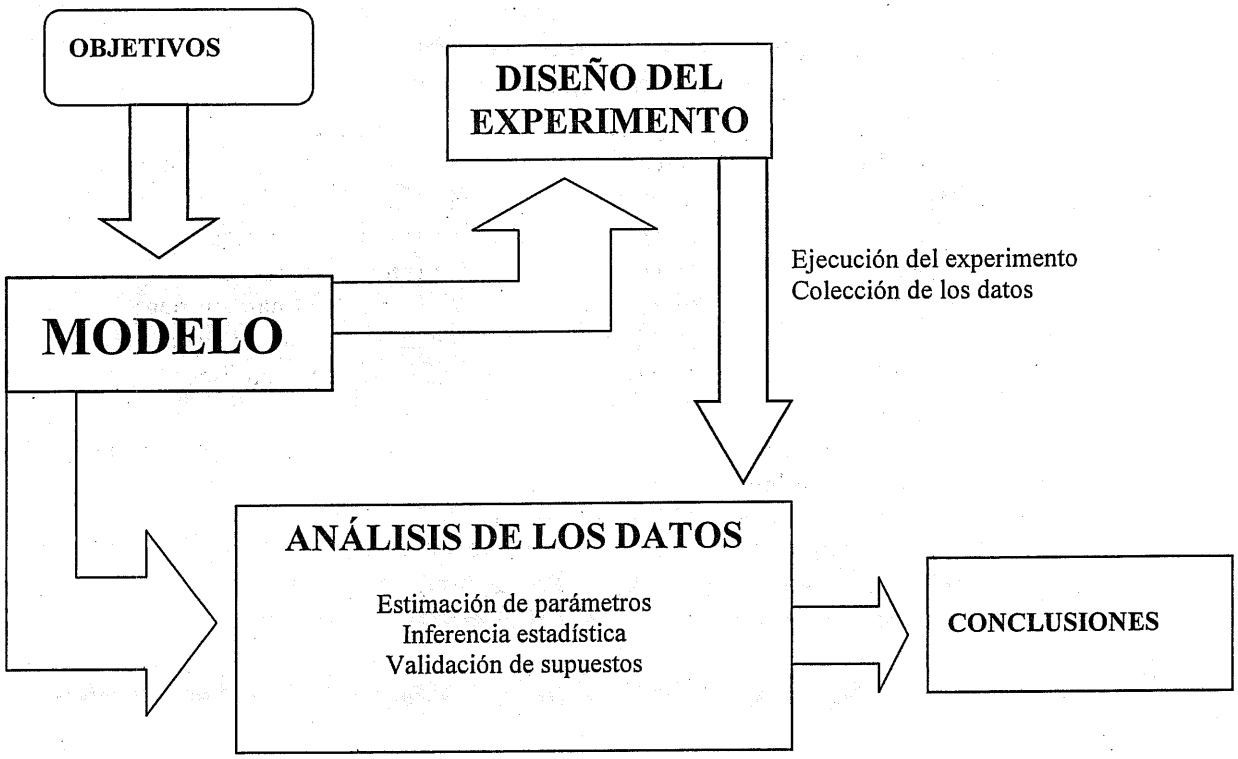


Figura 1.8. La complejidad del modelo y los parámetros a estimar son determinados por los objetivos del proyecto y al mismo tiempo son los determinantes de las fuentes de variación planificadas a ser consideradas en el diseño del experimento y en el análisis de los datos.

Asumiendo que se ha elegido el modelo apropiado para representar el comportamiento de la variable respuesta como función de las variables explicativas, entonces se pueden extraer ciertas conclusiones acerca de las hipótesis y extenderlas a la población de referencia, siempre y cuando la metodología del estudio incorpore la aleatorización y la repetición correspondientes al modelo y el análisis propuestos. Para que se pueda valuar la inseguridad de las conclusiones basadas en los datos mediante el cálculo de probabilidades, se deben cumplir los *supuestos* de cada método de análisis, que serán presentados en detalle en cada uno de los capítulos siguientes. Finalmente, el

grado de cumplimiento de los objetivos del estudio dependerá de la definición y correspondencia entre la *población de referencia*, el *modelo*, el *muestreo* o *diseño*, el *análisis* y las *conclusiones*.

11. Estimación

¿Qué es lo que puede decirse acerca de la diferencia verdadera entre dos tratamientos? El análisis de los resultados no puede dar el valor exacto de la diferencia verdadera (que no se conoce), y tampoco es posible determinar dos límites dentro de los cuales tengamos certeza de encontrar el valor exacto. En cambio, se puede hacer una afirmación para una probabilidad elegida, por ejemplo 0.95 o 0.99, de que la diferencia verdadera se encuentra entre dos límites calculados con los datos. Es decir, se puede establecer la probabilidad de que sea correcta la afirmación que establece límites para la diferencia verdadera.

¿Será esta información suficientemente precisa para que haga posible tomar decisiones de importancia práctica? En general, sí, a menos que los datos hayan sido insuficientes en cantidad o en calidad; por ejemplo, supongamos que la aplicación de una nueva tecnología de cultivo resultara beneficiosa sólo en el caso de que aumente el rendimiento en 80 kg (porque en ese caso el beneficio supera al costo diferencial de aplicar la nueva práctica). Supongamos además que se realiza un experimento y se obtienen límites de confianza para el efecto de la práctica de 130 y 345 kg; entonces se puede recomendar su aplicación con la tranquilidad de que resultará efectiva. De igual forma, si los límites resultaran -15 y 45 kg, se decidiría no introducir la nueva práctica. En cambio, si los límites resultaran de 10 y 165 kg, no se puede hacer ninguna recomendación sin gran riesgo de equivocarse. Para tomar la decisión correcta en esta última situación se debe realizar un experimento que acorte la distancia entre los límites de confianza.

12. Las pruebas de hipótesis estadísticas

Al poner a prueba una hipótesis estadística, el interés está centrado en someter a prueba una suposición de que la diferencia verdadera en las respuestas tiene algún valor específico, siendo cero el caso más común. El cero (que indica que no hay diferencias entre tratamientos) es la hipótesis nula que dice "nada interesante está ocurriendo". Si los datos aportaran evidencias para refutarla estamos en condiciones de concluir que hay un efecto significativo, equivalente a decir que los tratamientos difieren en sus efectos. La dificultad surge porque, como resultado de su variabilidad, los datos nunca están exactamente de acuerdo con las hipótesis y, entonces, el problema reside en decidir si la discrepancia entre los datos y la hipótesis va a ser atribuida a esas variaciones o al hecho de que la hipótesis sea falsa.

Las reglas para decidir si una hipótesis se va a considerar falsa están construidas para satisfacer dos condiciones claramente deseables: (i) las hipótesis que sean ciertas serán rechazadas sólo ocasionalmente y la probabilidad de que esto ocurra puede ser fijada previamente y, (ii) las hipótesis que sean falsas serán rechazadas tan frecuentemente como sea posible.

Una propiedad útil de este procedimiento es su influencia moderadora sobre aquel experimentador que tiende a extraer conclusiones fundamentadas en datos insuficientes y pretende convencer a los demás acerca de un efecto marcado de un tratamiento. El

riesgo de rechazar erróneamente la hipótesis de *no efecto* cuando es correcta (probabilidad de error de tipo I), se denomina ***nivel de significación*** de la prueba. Las probabilidades más frecuentemente utilizadas para este riesgo son 0.05 y 0.01 pero estos niveles son únicamente convenciones útiles y pueden usarse probabilidades más bajas si fueran muy serias las consecuencias de un rechazo erróneo de la hipótesis.

El nivel de significación de una prueba pierde importancia cuando los datos de un experimento no aportan suficiente evidencia para rechazar la hipótesis nula. Por ejemplo, si el resultado de un estudio fuera "no se detectaron diferencias en la función hepática entre personas tratadas con la dosis más alta de un nuevo analgésico para calmar dolores de cabeza y aquellas que recibieron un placebo", ¿eso significa que el analgésico es inocuo para el funcionamiento del hígado, aún en su dosis más alta?

¿Se puede suponer que la hipótesis nula es verdadera e interpretar que el efecto es cero? Evidentemente no. Cuando un investigador comunica que no encontró diferencias significativas, necesitamos que nos informe qué potencia tenía su prueba, es decir qué probabilidad de evitar un error de tipo II (aceptar una hipótesis nula falsa).

Algunas definiciones:

Hipótesis nula (H_0): Hipótesis que generalmente indica que no existen diferencias.

Hipótesis alternativa (H_1): Hipótesis que indica que existen diferencias.

Error Tipo I: Rechazo de una hipótesis nula verdadera.

Error Tipo II: Aceptación de una hipótesis nula falsa.

Alfa: Probabilidad de cometer un Error Tipo I.

Beta: Probabilidad de cometer un Error Tipo II.

Potencia: Capacidad de una prueba estadística para rechazar la hipótesis nula cuando no es verdadera. Probabilidad de no cometer error Tipo II.

13. EXPERIMENTOS MANIPULATIVOS Y MENSURATIVOS: alcances de las conclusiones

Los análisis estadísticos para experimentos manipulativos y mensurativos generalmente son los mismos, la principal diferencia entre ambos surge al atribuir causas a las respuestas observadas (Hulbert 1984). En los experimentos manipulativos, uno puede asociar causas con tratamientos, es así que la variación en la variable respuesta se "particiona" en efecto de la variable independiente (tratamientos) y efecto de todos los otros factores o error experimental (Figura 1.5).

En cambio, los experimentos mensurativos no permiten separar fácilmente al factor bajo estudio de todas las otras causas (o fuentes) de variación. Por ejemplo, en un estudio de sociología rural, se deseaba conocer los efectos de la "Flexibilización laboral" aplicada durante la década del 90, sobre el número de trabajadores empleados que acceden a los beneficios (e.g., obra social, seguro de trabajo, jubilación, indemnización, aguinaldo). Para ello, se relevó el número de trabajadores con descuento jubilatorio en cinco departamentos de la provincia de Córdoba, cuya principal actividad productiva es el agro. Los relevamientos se realizaron en dos momentos (tratamientos): antes de la aplicación de la "Flexibilización laboral" (año 1990) y después de su aplicación (año 2003). Una vez obtenidos los datos, el análisis estadístico podrá mostrar si en el año 1990 hubo más o menos trabajadores con descuento jubilatorio que en el año 2003, y asociar una probabilidad de error a la conclusión elegida. Sin embargo, en caso de detectar una disminución en los trabajadores con descuento jubilatorio en el año 2003, esta diferencia no puede ser atribuida exclusivamente a la flexibilización laboral. Esto se debe a que otros factores (globalización, convertibilidad, etc.) que afectan a la

variable respuesta también han cambiado durante este período. Es decir, no se puede separar el efecto de la aplicación de la flexibilización laboral de estos otros factores. La causalidad de la respuesta observada no puede ser atribuida exclusivamente al factor de interés (la flexibilización laboral). En cambio, si se puede separar el efecto de los tratamientos y los otros factores que "covarían" con ellos, de la variación observada entre los distintos partidos de la provincia, que son incluidos en el error experimental. Por lo tanto, la partición de la variabilidad en un experimento mensurativo es entre el efecto de los tratamientos junto con otros posibles factores que varíen conjuntamente con ellos y el error experimental (Figura 1.9.).

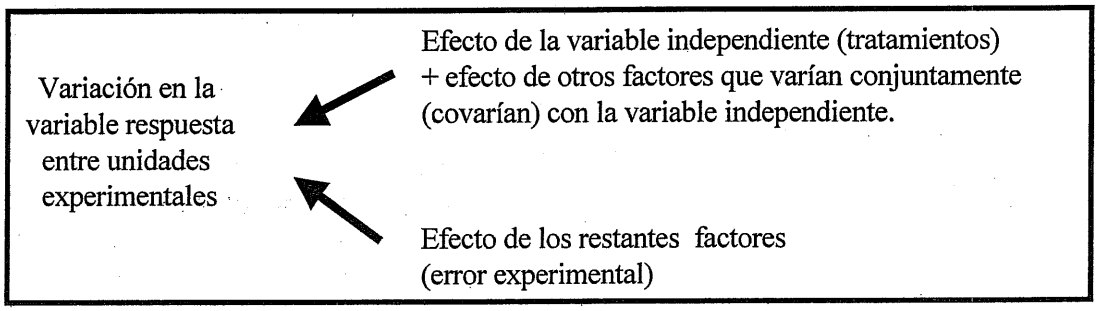

Variación en la variable respuesta entre unidades experimentales
Efecto de la variable independiente (tratamientos) + efecto de otros factores que varían conjuntamente (covarían) con la variable independiente.
Efecto de los restantes factores (error experimental)

Figura 1.9. Determinantes de la variación entre distintas unidades en la variable respuesta en experimentos mensurativos cuando existen factores que covarían con la variable independiente.

Por lo tanto, es importante diseñar el experimento mensurativo de manera de poder separar la mayor cantidad de efectos de otros factores que puedan confundirse con el efecto de los tratamientos. Sin embargo, a diferencia de los experimentos manipulativos, nunca se podrá separar completamente el efecto de los factores de interés de todos los factores restantes. Esto no significa que los experimentos mensurativos sean "malos" experimentos y los experimentos manipulativos sean "buenos". Los experimentos manipulativos permiten atribuir la proporción de la variación total que se debe exclusivamente al factor bajo estudio, pero muchas veces carecen de realismo. En cambio, los experimentos mensurativos permiten estudiar los sistemas de interés directamente. Es por esto que se necesitan de los dos tipos de experimentos para avanzar en el conocimiento, y que generalmente se incluyen diferentes combinaciones de experimentos mensurativos y manipulativos en los programas científicos. Más aún, muchas veces no es ético realizar experimentos manipulativos (e.g. con personas) y gran parte de nuestro conocimiento se basa en experimentos mensurativos. Lo importante es que, ya sea que tengamos que basarnos solamente en experimentos mensurativos, no utilicemos análisis estadísticos incorrectamente para extraer conclusiones más allá de las que el tipo de experimento permite.

BIBLIOGRAFÍA CITADA

Eberhardt, L. L. & Thomas, J. M. 1991. Designing environmental field studies. *Ecological Monographs* 61:53-73.

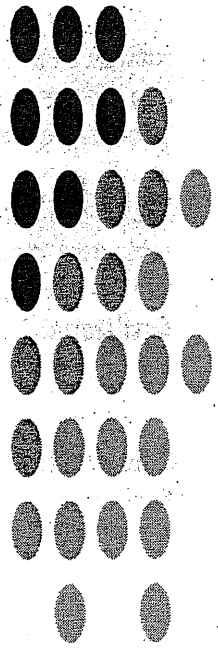
Farji-Brener, A. 2003. Uso correcto, parcial e incorrecto de los términos "hipótesis" y "predicciones" en Ecología. *Ecología Austral* 13:223-227.

Hulbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54:187-211.

Capítulo 2
El Análisis de la Varianza

Una paradoja: mediante el estudio de la variabilidad se comparan los promedios de las respuestas a los tratamientos

Una de las técnicas más poderosas y versátiles de la estadística



ANÁLISIS DE LA VARIANZA

El análisis de la varianza (o Anova: **An**alysis **o**f **va**riance), es un método que permite comparar varias medias en diversas situaciones. Este análisis se basa en la partición de la variación de las respuestas observadas, descomponiéndola en fuentes de variación independientes, como ser tratamientos, bloques, covariables y error experimental. El ANOVA está muy relacionado, por lo tanto, al diseño de experimentos y a los modelos estadísticos.

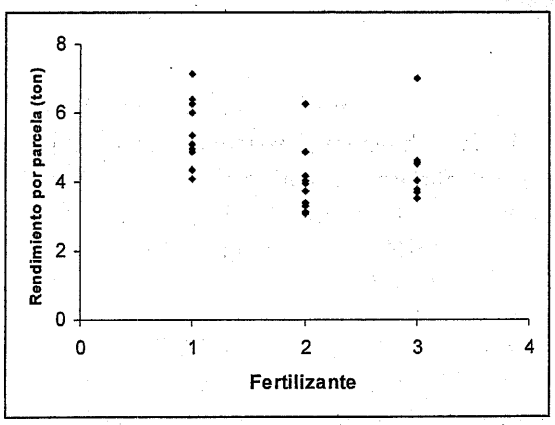
Si bien el ANOVA es una técnica de gran versatilidad, que permite analizar datos asociados con modelos de muy variada complejidad, para su presentación vamos a considerar el problema más simple que es la comparación de tres medias poblacionales. En los sucesivos capítulos abordaremos la generalización para modelos más complejos que comprenden varios factores. Para el caso particular de sólo dos poblaciones, el ANOVA es exactamente equivalente a la prueba t de comparación de dos medias poblacionales a partir de muestras independientes, aunque a primera vista aparenta ser un análisis completamente diferente, como se podrá apreciar a continuación.

Si se desea comparar la efectividad de tres fertilizantes (1, 2 y 3), se puede aplicar cada fertilizante a 10 parcelas cultivadas. Una vez transcurrida la estación de crecimiento, se realizará la cosecha de las 30 parcelas y se medirá el rendimiento del cultivo en cada una de ellas. De esta manera se podrían obtener tres grupos con 10 datos cada uno (como los de la Tabla 1). Queremos saber si existen diferencias interesantes entre los rendimientos promedio de esos grupos. Cuando los datos se representan en un gráfico, parece que los fertilizantes difieren en la magnitud del rendimiento producido, pero también existe considerable variación entre parcelas que recibieron el mismo fertilizante (Figura 1). Si bien el fertilizante 1 aparenta tener un mayor rendimiento promedio, algunas parcelas tratadas con el fertilizante 1 rindieron menos que otras tratadas con los fertilizantes 2 o 3.

Tabla 1. Resultados de rendimiento del cultivo en un experimento en el que se prueban tres fertilizantes en 30 lotes homogéneos. Cada fertilizante fue asignado al azar a 10 lotes.

FERTILIZANTE	Rendimiento en ton de las 10 parcelas asignadas a cada fertilizante									
1	6.27	5.36	6.39	4.85	5.99	7.14	5.08	4.07	4.35	4.95
2	3.07	3.29	4.04	4.19	3.41	3.75	4.87	3.94	6.28	3.15
3	4.04	3.79	4.56	4.55	4.53	3.53	3.71	7.00	4.61	4.55

Figura 1: Rendimiento de las 10 parcelas asignadas a cada uno de los fertilizantes.



Ahora necesitamos comparar estos tres grupos para descubrir si estas diferencias son estadísticamente significativas. En la comparación de dos medias poblacionales con la prueba t de Student el primer paso consistía en el cálculo de la diferencia entre los promedios muestrales. En cambio en el ANOVA no se calculan directamente las diferencias entre los promedios sino que se observa la variabilidad de los datos. En principio esto parece un poco contra-intuitivo: ¿vamos a contestar preguntas acerca de los rendimientos medios de los tres grupos a través del análisis de la variación en los datos? Veamos de qué manera funciona esto...

¿QUÉ OCURRE CUANDO CALCULAMOS UNA VARIANZA?

La variabilidad de un conjunto de datos cuantifica la dispersión de los puntos alrededor del promedio. Para calcular la varianza, primero se calcula el promedio y luego se obtiene la desviación de cada punto con respecto a ese promedio. Las desviaciones calculadas serán algunas positivas y otras negativas de manera que su suma es cero, esto es siempre así sin importar el tamaño de la muestra (conjunto de datos) ni la variabilidad que presenten esos datos. En consecuencia, la simple suma de desvíos no sirve como medida de variabilidad. En cambio, si las desviaciones son elevadas al cuadrado antes de sumarlas, la suma obtenida es una medida de variabilidad muy útil, que se incrementa cuando la dispersión de los datos es mayor. Esta cantidad se denomina ***suma de cuadrados (SC)*** y es central en este análisis. En la figura 2 se pueden observar las desviaciones con respecto al promedio de los datos de fertilización, sin distinguir por el momento entre grupos de parcelas tratadas con los diferentes fertilizantes.

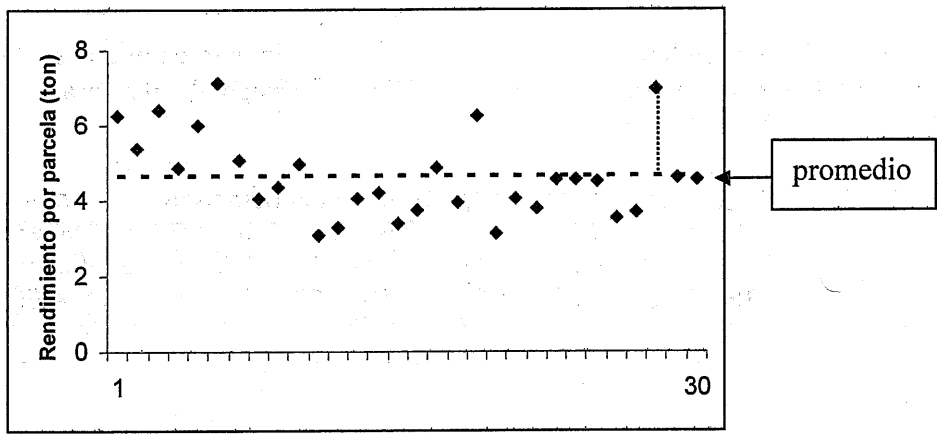


Figura 2: Rendimiento por parcela. Las distancias verticales desde cada punto al promedio (sólo una está representada como ejemplo) son las desviaciones. En el eje de abscisas se representan las 30 parcelas sin diferenciar por tratamientos.

La SC, sin embargo, estará claramente influenciada por el número de observaciones de cada conjunto de datos (a más puntos, con la misma dispersión total, mayor SC) en consecuencia no puede ser utilizada como una medida comparativa entre grupos con distinto número de observaciones. Pero esta medida se transforma en una varianza al dividirla por n-1, donde n representa el número de puntos en el conjunto de datos. La varianza es entonces una medida de la variabilidad que toma en cuenta el tamaño de la muestra.

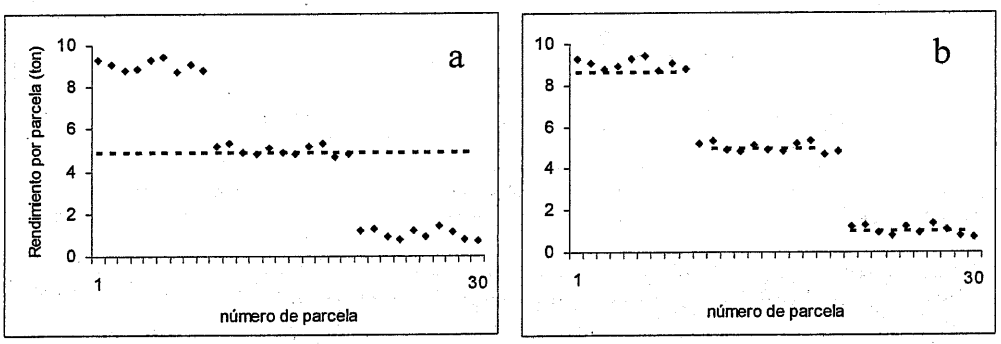
¿POR QUÉ USAMOS n-1 EN LUGAR DE n?

Si queremos calcular un promedio de las desviaciones cuadradas respecto a la media, ¿por qué no dividir por n? La razón es que no tenemos n piezas independientes de información acerca de la varianza. El primer paso consistió en calcular la media a partir de las n piezas independientes de datos que hemos coleccionado. El segundo paso consiste en calcular una varianza con referencia a esa media. Una vez que se calculan n-1 desviaciones, ya se conoce el valor de la última, dado que deben sumar cero por definición. En consecuencia, sólo contamos con n-1 piezas independientes de información acerca de la variabilidad respecto al promedio. Es por esto que tiene más sentido dividir la SC por n-1 para obtener el promedio de las desviaciones cuadradas, que se denomina ***cuadrado medio (CM)***. El número de piezas independientes de información que contribuyen al cálculo de un estadístico se denomina en general ***grados de libertad***.

PARTICIÓN DE LA VARIABILIDAD

En relación a nuestra pregunta inicial:¿ qué causa la variación entre los rendimientos de las 30 parcelas? Podemos afirmar que probablemente estén comprometidos numerosos factores, como las diferencias en el nivel de nutrientes entre los suelos de las parcelas, las diferencias en los contenidos de humedad, muchos otros factores bióticos y abióticos, y también los fertilizantes distintos que hemos aplicado en las parcelas. En esta última fuente de variación es en la que estamos particularmente interesados, por eso vamos a repartir la variabilidad en dos partes: la que se debe a la aplicación de los diferentes fertilizantes y aquella debida a todos los otros factores (no controlados en el experimento) en conjunto. Para ilustrar el principio que subyace a esta partición, primero vamos a considerar dos conjuntos de datos ficticios muy extremos y luego volveremos a nuestro ejemplo.

Figura 3: Ejemplo ficticio 1, las desviaciones con respecto a la media general (distancias verticales a la línea de puntos en la figura **a**) son mayormente explicadas por las diferencias entre los promedios de los tratamientos, como se desprende de la figura **b** , donde se ve que al ajustar por los promedios de los grupos las desviaciones resultan muy pequeñas.



Si hubiera muy poca variación entre los rendimientos debida a otros factores y casi toda la variación fuera causada por los fertilizantes, entonces los datos presentarían un patrón como el de la figura 3a. Allí vemos que hay considerable dispersión de los puntos con respecto a la media gene-

ral (línea punteada horizontal). Por otra parte, si calculamos un promedio para cada uno de los grupos, las desviaciones con respecto a ellos resultan muy pequeñas (Figura 3b). En otras palabras, las diferencias entre los promedios de los grupos explican la mayor parte de la variabilidad de los datos. Esto es así porque las medias de los distintos tratamientos son muy diferentes.

Consideremos ahora otro ejemplo extremo completamente opuesto al anterior, en el que los tres fertilizantes son idénticos (Figura 4a). En este caso, al considerar la dispersión con respecto a los promedios de los grupos (Figura 4b), prácticamente se encuentra igual variabilidad que con respecto a la media general.

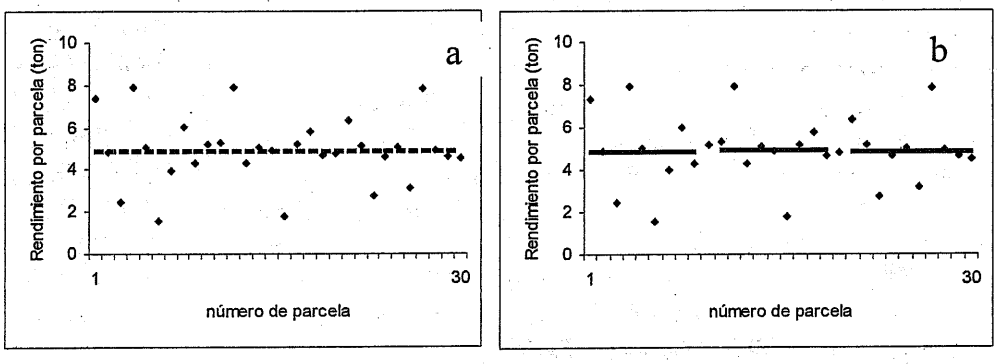


Figura 4: Ejemplo ficticio 2, las desviaciones con respecto a la media general (distancias verticales a la línea de puntos en la figura **a**) prácticamente no son explicadas por las diferencias entre los promedios de los tratamientos, como se desprende de la figura **b** , donde se ve que al ajustar por los promedios de los grupos las desviaciones permanecen casi iguales.

La cantidad de variabilidad explicada en cada uno de estos casos extremos, al ajustar por los promedios de los grupos, se puede comparar mirando la dispersión de las medias de los grupos con respecto a la media general (Figuras 5a y 5b), que en el primer ejemplo es muy grande (los fertilizantes tienen efectos muy diferentes) y en el segundo muy pequeña.

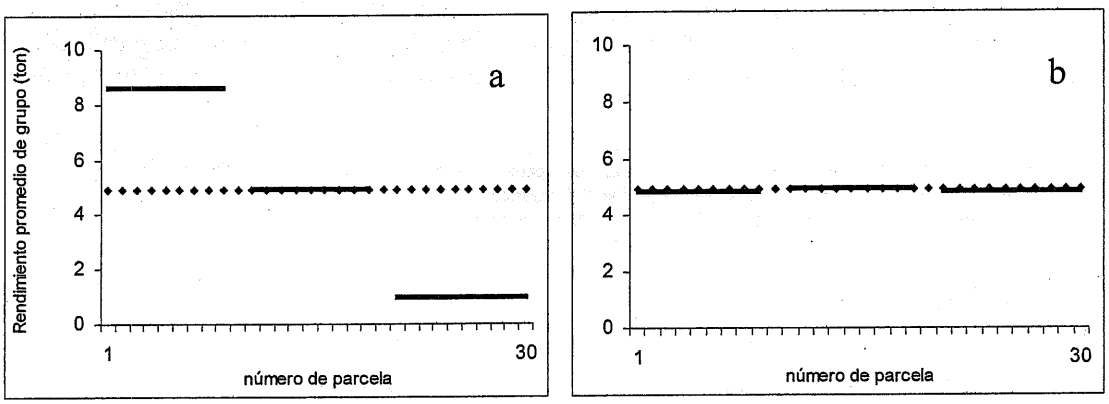


Figura 5: Desviaciones de las medias de los grupos con respecto a la media general (línea de puntos) para los ejemplos ficticios de las figuras 3 y 4 . En el primer ejemplo (**a**) las desviaciones de los promedios a la media general son considerables, en cambio en el segundo (**b**) , esas desviaciones son muy pequeñas.

El conjunto de datos de la Tabla 1 representa una situación intermedia entre los dos ejemplos ficticios, donde ya no es tan obvio si los fertilizantes tuvieron un efecto diferencial sobre el rendimiento del cultivo. En este caso, para poder decidir si al ajustar por los promedios de los tratamientos se ha explicado una porción significativa de la variación total, se requiere un procedi-

miento analítico. En este punto va a ser muy útil definir las tres medidas de variabilidad a las que nos referimos hasta ahora:

SC_{TOT}= suma de cuadrados total. Suma de los cuadrados de las desviaciones de los datos con respecto al promedio general. Es una medida de la variabilidad total de los datos.

SC_E= Suma de cuadrados del error. Suma de los cuadrados de las desviaciones de los datos con respecto a los tres promedios grupales respectivos. Es la variación entre parcelas que recibieron el mismo tratamiento, es decir la variación observada entre repeticiones que recibieron el mismo tratamiento.

SC_{TRAT}= Suma de cuadrados de tratamientos (fertilizantes en nuestro ejemplo). Suma de los cuadrados de las desviaciones de los promedios grupales con respecto al promedio general. Es la variación entre parcelas que recibieron distintos fertilizantes.

Se puede demostrar que estas tres cantidades cumplen la siguiente relación simple, que se denomina ***partición de las sumas de cuadrados***:

$SC_{TOT} = SC_{TRAT}$ (entre grupos) + SC_E (dentro de grupos)

De tal forma la variabilidad total ha sido dividida en dos componentes, la que se debe a diferencias entre parcelas que recibieron distintos tratamientos y aquella que ocurre entre parcelas que recibieron el mismo tratamiento.

PARTICIÓN DE LOS GRADOS DE LIBERTAD

Cada una de las sumas de cuadrados fue calculada utilizando un determinado número de piezas independientes de información. La SC_{TOT}, que comprende los desvíos de las **n** observaciones con respecto a la media general, tiene **n-1** desvíos independientes, en nuestro ejemplo $GL_{TOT} = 30 - 1 = 29$. La SC_{TRAT}, que corresponde a los desvíos de las **t** medias de tratamientos con respecto a la media general, tiene **t-1** desvíos independientes, en nuestro ejemplo $GL_{TRAT} = 3 - 1 = 2$. Finalmente la SC_E, que mide la variación de las n observaciones respecto a los promedios de los tratamientos consta de **t.(r-1)** desvíos independientes, en nuestro ejemplo $GL_E = 3\ (10 - 1) = 27$. Donde r representa el número de repeticiones dentro de cada tratamiento, de manera que: **n = t. r**

Al igual que las sumas de cuadrados, los grados de libertad son aditivos, es decir:

$$GL_{TOT} = GL_{TRAT} + GL_E = (t - 1) + t(r - 1) = tr - 1 = n-1$$

CUADRADOS MEDIOS

Los cuadrados medios (CM) constituyen una medida de variabilidad por grado de libertad y son estimadores de varianzas. A diferencia de las sumas de cuadrados y los grados de libertad, los CM no son aditivos.

$$CM_{TOT} = SC_{TOT}/GL_{TOT}$$

$$CM_{TRAT} = SC_{TRAT}/GL_{TRAT}$$

$$CM_E = SC_E/GL_E$$

Ahora sí estamos en condiciones de preguntar: al ajustar por las medias de tratamientos, ¿hemos explicado una porción significativa de la varianza total? Pero antes de contestar esta pregunta presentaremos una distribución de probabilidad que nos permitirá medir la incertidumbre de nuestras conclusiones.

La distribución F de Snedecor

Supongamos que se extrae una muestra aleatoria de n_1 observaciones de una distribución normal con varianza σ_1^2, se extrae una segunda muestra al azar de n_2 observaciones de una segunda distribución normal con varianza σ^2_2, y se calculan las estimaciones S^2_1 y S^2_2 de las dos varianzas poblacionales, teniendo ν_1 y ν_2 grados de libertad. Entonces S^2_1/σ_1^2 se distribuye como $\chi^2_{\nu 1}/\nu_1$, y S^2_2/σ^2_2 se distribuye como $\chi^2_{\nu 2}/\nu_2$.
Ahora, el cociente $(\chi^2_{\nu 1}/\nu_1)/(\chi^2_{\nu 2}/\nu_2)$ tiene una distribución F con ν_1 y ν_2 grados de libertad. Así

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \approx F_{\nu 1,\nu 2} \quad o \quad equivalente, \quad \frac{S_1^2}{S_2^2} \approx \frac{\sigma_1^2}{\sigma_2^2} \quad F_{\nu 1,\nu 2}$$

La distribución F puede así ser empleada para probar la hipótesis concerniente a σ_1^2/σ^2_2, el cociente de varianzas poblacionales de dos distribuciones normales. En particular, la hipótesis nula de que las varianzas son iguales ($\sigma_1^2=\sigma^2_2$) puede ser probada refiriendo el cociente de las varianzas muestrales S^2_1/S^2_2 directamente a la tabla F.

EL COCIENTE *F* Y LA PRUEBA DE HIPÓTESIS

Si ninguno de los fertilizantes hubiera tenido una influencia diferente sobre el rendimiento, la variación entre parcelas que recibieron distintos fertilizantes hubiera resultado muy semejante a la variación entre parcelas que recibieron el mismo. Es decir que CM_{TRAT} y CM_E, estimarían la misma varianza poblacional en la hipótesis de que las *t* muestras provengan de la misma población. Esto se puede expresar como un cociente de cuadrados medios, es decir:

$$\boldsymbol{CM_{TRAT}/\ CM_E = 1}$$

Sin embargo, aún si los fertilizantes tuvieran idéntico efecto, sería bastante poco probable que el cociente arrojara como resultado exactamente el valor 1, sino que por azar podría alcanzar un amplio rango de valores. Se necesita una distribución de probabilidad asociada a este cociente, que nos permita cuantificar qué tan raro será el resultado obtenido si verdaderamente los tres fertilizantes no presentan efectos diferenciales. Una distribución que represente la probabilidad de todos los posibles resultados del cociente cuando sea verdadera la hipótesis nula que postula que no hay efecto de tratamientos.

La distribución muestral del cociente de dos estimaciones independientes de la varianza de una población **normal** es una distribución ***F*** con los grados de libertad correspondientes al numerador y denominador respectivamente, por lo tanto se puede contrastar la hipótesis de interés utilizando esa distribución de probabilidad.

Si sobre la base de este contraste se rechaza la hipótesis de que CM_{TRAT} y CM_E estimen la misma varianza, se puede rechazar la hipótesis de que las *t* medias provengan de una misma población. Partiendo del supuesto que las muestras provengan de poblaciones con la misma varianza, este

rechazo implica que las medias poblacionales son distintas, de modo que con un único contraste se somete a prueba la igualdad de las t medias. La hipótesis nula en este análisis postula que todas las medias de rendimiento de las parcelas tratadas con distintos fertilizantes son iguales y la alternativa que al menos alguna difiere:

H_o: $\mu_1 = \mu_2 = \mu_3$

H_1: al menos una μ_i es diferente

Los resultados de un ANOVA se suelen representar en una tabla como la siguiente:

Fuente de variación	G.L.	SC	CM	F
Entre grupos **Tratamientos**	t-1	SC_{TRAT}	$CM_{TRAT} = SC_{TRAT}/(t\text{-}1)$	CM_{TRAT}/CM_E
Dentro de grupos **Error**	$(r\text{-}1)t$	SC_E	$CM_E = SC_E/t(r\text{-}1)$	
Total	n-1	SC_{TOT}		

F es el estadístico que se usa para realizar el contraste de la hipótesis de medias iguales. La región crítica para dicho contraste es $F > F_{t-1,(r-1)t}$, en otras palabras: la regla de decisión para que la hipótesis nula sea rechazada es que la probabilidad (el P-valor) asociada al cociente de cuadrados medios (estadístico F) resulte menor que alfa.

Recordemos que alfa, denominado ***nivel de significación*** de la prueba, es la probabilidad de cometer un error de tipo I: rechazar la hipótesis nula cuando en realidad los tratamientos no difieren en su efecto. Es decir que cuando elegimos un nivel de significación del 5%, por ejemplo, estamos decidiendo que vamos a considerar al cociente de cuadrados medios significativamente mayor que 1, siempre que arroje un valor que tiene probabilidad menor o igual al 5% de ocurrir cuando la hipótesis nula es correcta. En consecuencia, el valor P es una medida de la fuerza de la evidencia que aportan los datos en contra de la hipótesis nula. Sólo cuando valor P es menor que alfa, consideramos que hay suficiente evidencia para afirmar que los tratamientos producen efectos diferentes.

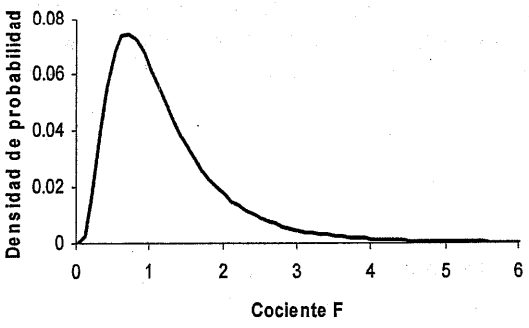
Los resultados del ANOVA para los datos de la Tabla 1 se presentan en la siguiente tabla:

Fuente de Variación	SC	GL	CM	F	Valor P
FERTILIZANTE	10.82	2	5.41	5.7	0.0086
Error	25.62	27	0.95		
Total	36.44	29			

Los cálculos detallados que conducen a los resultados presentados en la tabla se muestran en la Tabla 2. La presentación formal de la partición de sumas de cuadrados con fórmulas matemáticas se presenta en el Cuadro 1: Un poco de notación. Recordemos que la hipótesis nula en este análisis postula que todas las medias de rendimiento de las parcelas tratadas con distintos fertilizantes son iguales y la alternativa que al menos alguna difiere

H_o: $\mu_1 = \mu_2 = \mu_3$

H_1: al menos una μ_i es diferente



El valor P asociado al cociente F que compara el CM_{TRAT} con el CM_E es 0.0086, es decir que la probabilidad de obtener un cociente mayor o igual a 5.7 si la hipótesis nula es verdadera es 0.0086, suficientemente pequeña como para concluir que estos fertilizantes difieren en su eficacia sobre el rendimiento del cultivo. En consecuencia, se rechaza la hipótesis nula y se concluye que al menos uno de los fertilizantes difiere de los demás en su eficacia.

Con estos ejemplos hemos visto cómo se comparan medias poblacionales estudiando la variabilidad de los datos, en un intento por aclarar la paradoja que encierra el nombre de la técnica. En el ANOVA se comparan medias, no varianzas como su nombre podría hacer suponer: se comparan medias de los grupos o estratos originados por los factores de clasificación estudiados. El análisis de la varianza en su concepción más amplia particiona la suma de los cuadrados de los desvíos llamada suma de cuadrados total (SC_{TOT}) en partes, cada una de las cuales se atribuye a una de las variables independientes en el modelo (SC_{TRAT}), más un residuo que se asocia con el error aleatorio, llamado suma de cuadrados del error (SC_E). En consecuencia, las técnicas del análisis de la varianza, basadas en la partición de la variabilidad (suma de cuadrados) de la respuesta estudiada, sirven de base para el análisis de distribuciones de datos generados en diseños más complejos, algunos de los cuales serán presentados en los siguientes capítulos. En el ejemplo sencillo de la Tabla 1, el modelo presenta solo una variable independiente (Fertilizantes) y el análisis por lo tanto se denomina ***ANOVA de una vía o de un solo factor.***

Recordemos entonces que...

Se utiliza el nombre de ***análisis de la varianza*** ya que el elemento básico del análisis estadístico será precisamente el estudio de la variabilidad. Teóricamente es posible dividir la variabilidad de la variable respuesta que se estudia en dos partes:

- La originada por los tratamientos bajo estudio;

- La producida por los restantes factores que entran en juego, conocidos o no, controlables o no, que se denomina error experimental.

Si mediante los contrastes estadísticos adecuados, la variación producida por los tratamientos resulta significativamente mayor que la producida por el error experimental, podemos aceptar la hipótesis de que los distintos tratamientos actúan de forma distinta sobre la variable respuesta. Lo cual es equivalente a rechazar la hipótesis nula de "nada interesante está ocurriendo".

En esta técnica, la hipótesis alternativa es múltiple, puesto que engloba diferentes hipótesis, todas ellas referidas al hecho de que alguna de las medias de los estratos sea diferente de las demás. Cuando se alcanza la significación en alguno de los factores estudiados, es lógico preguntarse cuál de las hipótesis alternativas es la que origina la significación. Esto es equivalente a preguntarse qué media o medias son diferentes entre sí. Estos enigmas se resuelven mediante procedimientos "a posteriori" para comparar más de dos medias. Son procedimientos denominados de ***comparaciones múltiples***, que funcionan eficientemente, siempre que previamente el factor en estudio haya resultado significativo en el ANOVA. Antes de presentarlos vamos a exponer el modelo lineal aditivo y los supuestos en los que se fundamenta el ANOVA.

FERT	Y_{ik}	$\overline{Y}_{..}$	$\overline{Y}_{i.}$	$\overline{Y}_{ik} - \overline{Y}_{..}$	$Y_{ik} - \overline{Y}_{i.}$	$\overline{Y}_{i.} - \overline{Y}_{..}$	$(Y_{ik} - Y..)^2$	$(Y_{ik} - \overline{Y}_{i.})^2$	$(\overline{Y}_{i.} - \overline{Y}_{..})^2$
1	6.27	4.644	5.445	1.626	0.825	0.801	2.645	0.681	0.642
1	5.36	4.644	5.445	0.716	-0.085	0.801	0.513	0.007	0.642
1	6.39	4.644	5.445	1.746	0.945	0.801	3.050	0.893	0.642
1	4.85	4.644	5.445	0.206	-0.595	0.801	0.043	0.354	0.642
1	5.99	4.644	5.445	1.346	0.545	0.801	1.813	0.297	0.642
1	7.14	4.644	5.445	2.496	1.695	0.801	6.232	2.873	0.642
1	5.08	4.644	5.445	0.436	-0.365	0.801	0.190	0.133	0.642
1	4.07	4.644	5.445	-0.574	-1.375	0.801	0.329	1.891	0.642
1	4.35	4.644	5.445	-0.294	-1.095	0.801	0.086	1.199	0.642
1	4.95	4.644	5.445	0.306	-0.495	0.801	0.094	0.245	0.642
2	3.07	4.644	3.999	-1.574	-0.929	-0.645	2.476	0.863	0.416
2	3.29	4.644	3.999	-1.354	-0.709	-0.645	1.832	0.503	0.416
2	4.04	4.644	3.999	-0.604	0.041	-0.645	0.364	0.002	0.416
2	4.19	4.644	3.999	-0.454	0.191	-0.645	0.206	0.036	0.416
2	3.41	4.644	3.999	-1.234	-0.589	-0.645	1.522	0.347	0.416
2	3.75	4.644	3.999	-0.894	-0.249	-0.645	0.799	0.062	0.416
2	4.87	4.644	3.999	0.226	0.871	-0.645	0.051	0.759	0.416
2	3.94	4.644	3.999	-0.704	-0.059	-0.645	0.495	0.003	0.416
2	6.28	4.644	3.999	1.636	2.281	-0.645	2.678	5.203	0.416
2	3.15	4.644	3.999	-1.494	-0.849	-0.645	2.231	0.721	0.416
3	4.04	4.644	4.487	-0.604	-0.447	-0.157	0.364	0.200	0.025
3	3.79	4.644	4.487	-0.854	-0.697	-0.157	0.729	0.486	0.025
3	4.56	4.644	4.487	-0.084	0.073	-0.157	0.007	0.005	0.025
3	4.55	4.644	4.487	-0.094	0.063	-0.157	0.009	0.004	0.025
3	4.53	4.644	4.487	-0.114	0.043	-0.157	0.013	0.002	0.025
3	3.53	4.644	4.487	-1.114	-0.957	-0.157	1.240	0.916	0.025
3	3.71	4.644	4.487	-0.934	-0.777	-0.157	0.872	0.604	0.025
3	7	4.644	4.487	2.356	2.513	-0.157	5.552	6.315	0.025
3	4.61	4.644	4.487	-0.034	0.123	-0.157	0.001	0.015	0.025
3	4.55	4.644	4.487	-0.094	0.063	-0.157	0.009	0.004	0.025
SC							**36.445**	**25.622**	**10.823**
GL							29	27	2
							SC_{TOT}	**SC_E**	**SC_{TRAT}**

Tabla 2 : Contribución de cada observación a las sumas de cuadrados para los datos de la Tabla 1. **FERT:** tratamiento de fertilizante, Y_{ik}: rendimiento, $\overline{Y}_{..}$: media general, $\overline{Y}_{i.}$: media de cada grupo de parcelas que recibieron igual fertilizante. $\overline{Y}_{ik} - \overline{Y}_{..}$: desviación de cada observación a la media general, es la contribución de la observación a la variación total. $Y_{ik} - \overline{Y}_{i.}$: desviación de cada observación a la media del respectivo tratamiento de fertilización, es la contribución de cada observación al error experimental; $\overline{Y}_{i.} - \overline{Y}_{..}$: desviación de la media de cada tratamiento de fertilización a la media general, es la parte de la variación explicada por el tratamiento.

CUADRO 1: UN POCO DE NOTACIÓN

La Tabla muestra la notación simbólica de los resultados muestrales de una investigación.

Nivel de factor	Observaciones					TOTALES	MEDIAS
	1	...	k	...	r		
1	y_{11}	...	y_{1k}	...	y_{1r}	$y_{1.}$	$\bar{y}_{1.}$
.	.	...	.	...	.	.	
.	.	...	.	...	.	.	
i	y_{i1}	...	y_{ik}	...	y_{ir}	$y_{i.}$	$\bar{y}_{i.}$
.	.	...	.	...	.	.	.
.	.	...	.	...	.	.	.
t	y_{t1}	...	y_{tk}	...	y_{tr}	$y_{t.}$	$\bar{y}_{t.}$
						$y_{..}$	$\bar{y}_{..}$

El punto como subíndice indica que se suma sobre el subíndice que se reemplaza por el punto y la barra sobre una variable indica promedio.

El total de los valores de las observaciones para el tratamiento i-ésimo es:

$$y_{i.} = \sum_{k=1}^{r} y_{ik}$$

La media del tratamiento i-ésimo es:

$$\bar{y}_{i.} = \frac{\sum_{k=1}^{r} y_{ik}}{r} = \frac{y_{i.}}{r}$$

El total de los valores de todas las observaciones es:

$$y_{..} = \sum_{i=1}^{t} \sum_{k=1}^{r} y_{ik}$$

La media de todas las observaciones es:

$$\bar{y}_{..} = \frac{\sum_{i=1}^{t} \sum_{k=1}^{r} y_{ik}}{t.r} = \frac{y_{..}}{n}$$

La variación total de la y_{ik} sin usar ninguna información acerca de los tratamientos, se mide en términos de las desviaciones de las observaciones y_{ik} alrededor de la media $\bar{y}_{..}$, o sea:

$$y_{ik} - \bar{y}_{..}$$

La medida convencional de la variabilidad total es la suma de cuadrados de estos desvíos expresada como SC_{TOT} (suma de cuadrados totales)

$$SC_{TOT} = \sum_{i=1}^{t} \sum_{k=1}^{r} (y_{ik} - \bar{y}_{..})^2$$

CUADRO 1: CONTINUACIÓN

PARTICIÓN DE LA SUMA DE CUADRADOS TOTAL

Cuando utilizamos información acerca de los tratamientos, las desviaciones que reflejan la variación remanente de los datos son aquellas de cada observación y_{ik} alrededor de la respectiva media del tratamiento $\bar{y}_{i.}$

$y_{ik} - \bar{y}_{i.}$

Notemos, ahora, como podemos descomponer la desviación total $y_{ik} - \bar{y}_{..}$ en dos componentes:

$y_{ik} - \bar{y}_{..} = (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ik} - \bar{y}_{i.})$

⇩	⇩	⇩
Desviación Total	Desviación del tratamiento respecto a la media general	Desviación respecto a la media del tratamiento

Así la desviación total puede ser vista como la suma de dos componentes.

Cuando elevamos al cuadrado y luego realizamos las sumas para todas las observaciones:

$$\sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{..})^2 = \sum_{i=1}^{t}\sum_{k=1}^{r}\left[(\bar{y}_{i.} - \bar{y}_{..}) + (y_{ik} - \bar{y}_{i.})\right]^2$$

Desarrollando el cuadrado de la derecha, los términos de los productos cruzados se anulan por ser sumas de desviaciones con respecto a la media. Así:

$$\sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{..})^2 = r\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{i.})^2$$

$$\Downarrow \qquad \Downarrow \qquad \Downarrow$$

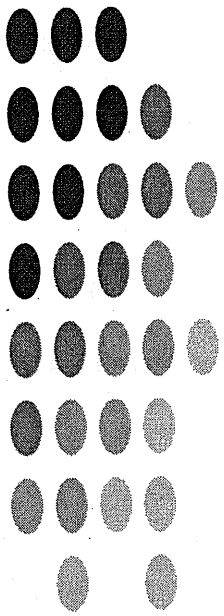
$$SC_{TOT} = SC_{TRAT} + SC_E$$

Esta partición de la suma de cuadrados total es idéntica a la que se presenta calculada para el problema de los fertilizantes en la Tabla 2.

Capítulo 3
El modelo lineal aditivo

El modelo vincula los objetivos con el diseño del experimento y el análisis de los datos

Las comparaciones posteriores al rechazo de la hipótesis nula



EL MODELO LINEAL ADITIVO

Una práctica común en todas las ciencias es intentar explicar fenómenos naturales mediante modelos. Por ejemplo, decimos que la tierra tiene forma esferoide y gira sobre su eje norte-sur. Esto nos permite explicar por qué desaparece un barco en el horizonte y la sucesión de los días y las noches, entre otros hechos observados y también nos permite predecir sucesos que aún no fueron observados, como la hora del amanecer en Buenos Aires un día de la semana que viene. Frecuentemente los modelos conceptuales pueden ser expresados matemáticamente, como el que representa la relación entre el volumen, el número de moles y la temperatura de un gas con la presión a la que está sometido: P.V=n.R.T.

En estadística, el modelo habitual para describir la naturaleza de una observación consta de una **media** más un **error**. Este es un *modelo lineal aditivo*. La media puede comprender sólo al parámetro μ o bien puede estar compuesta de una suma de parámetros. Los supuestos con respecto a los parámetros y los errores dependen de cada problema. Un supuesto mínimo es que los errores son aleatorios, esto significa que la población de las observaciones y se muestrea aleatoriamente. Esto hace que el modelo resulte estadístico en lugar de determinista.

El modelo lineal aditivo más simple es el que expresa que el valor k-ésimo de la variable y es una observación de la media μ, pero está sujeta a un error de muestreo ε_k (épsilon sub k):

$$y_k = \mu + \varepsilon_k$$

Los ε_k pertenecen a una población de ε con media cero y son función de la varianza de la población (σ^2) , ya que miden cuán lejos se encuentra cada y_k con respecto a la media de su población.

Para muestras de varias poblaciones con medias posiblemente diferentes (por ejemplo en respuesta a distintos tratamientos) pero con varianza común, la composición de una observación se expresa como:

$$y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

donde i = 1, 2, ..., t y k = 1, 2, .., r

El modelo explica a la repetición k-ésima del tratamiento i-ésimo como compuesta por una media general μ, más una componente τ_i (tau sub i) propia del respectivo tratamiento, más un elemento aleatorio de variación ε_{ik} propio de cada observación. El valor **t** indica el número de tratamientos y el valor **r** el número de repeticiones (observaciones en cada tratamiento).

Se supone que los ε_{ik} provienen de una sola población de ε con media $\mu = 0$ y varianza σ^2.

Tal como había sido presentado en el capítulo 1, el modelo expresa la respuesta observada a través de una suma de términos que representan las principales fuentes de variación planificadas más un término residual debido a las fuentes de variación aleatoria no planificada.

RESPUESTA = Constante + Efecto de tratamientos + error

$$y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

Más adelante veremos que simplemente adicionando nuevos términos al modelo se pueden representar otras fuentes planificadas de variación, como los bloques y/o los factores y sus interacciones en los experimentos factoriales.

ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO

Los parámetros del modelo son características poblacionales desconocidas y deben ser estimados a partir de los datos del experimento. Para estimar los parámetros de modelos lineales sencillos como los que abarca este curso se utilizan los estimadores de mínimos cuadrados, porque para estos casos resultan los mejores estimadores insesgados y de mínima varianza. A continuación se presentan los estimadores de mínimos cuadrados para μ, τ_i, y σ^2. Los efectos de tratamientos se definen como desviaciones de la media y por lo tanto: $\Sigma_i \tau_i = 0$

Parámetro	Estimador
μ	$\bar{y}_{..}$
τ_i	$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$
σ^2	**Cuadrado Medio del Error**

La componente de error e_{ik} tiene como estimador al respectivo residual

$$e_{ik} = y_{ik} - \bar{y}_{i.}$$

Para el problema de los fertilizantes presentado en el capítulo de Análisis de la Varianza, los estimadores de las medias de los tratamientos son:

$\bar{y}_{1.} = \frac{54,5}{10} = 5,45$ ← estimador de μ_1

$\bar{y}_{2.} = \frac{40,0}{10} = 4,00$ ← estimador de μ_2

$\bar{y}_{3.} = \frac{44,9}{10} = 4,49$ ← estimador de μ_3

El estimador de la media general es: $\bar{y}.. = \frac{139,4}{30} = 4,64$ ← estimador de μ

El estimador de la varianza es: $CM_E = 0.95$ ← estimador de σ^2

Los estimadores de los efectos de los tres tratamientos, son:

$\hat{\tau}_1 = 5,45 - 4,64 = 0,81$

$\hat{\tau}_2 = 4,00 - 4,64 = -0,64$

$\hat{\tau}_3 = 4,49 - 4,64 = -0,15$

LAS HIPÓTESIS Y EL MODELO

La hipótesis nula que estamos interesados en poner a prueba, postula que las medias poblacionales de la variable respuesta correspondientes a los distintos tratamientos son todas iguales y la alternativa postula que al menos alguna difiere. En el ejemplo de los tres fertilizantes, las hipótesis se expresan como:

H_0: $\mu_1 = \mu_2 = \mu_3$

H_1: al menos una μ_i es diferente

En términos de la hipótesis nula, todas las poblaciones tienen la misma media. En cambio, si H_0 no es verdadera, las poblaciones no tienen todas la misma media y se asume que difieren en una cantidad identificada mediante los términos τ_i (Figura 1), es decir que hay una diferencia lineal aditiva entre cada media poblacional y la media general de todas las poblaciones, o sea cada $\mu_i = \mu + \tau_i$

El modelo que incluye a los parámetros τ_i es válido para ambas situaciones (la que postula la hipótesis nula y la que postula la hipótesis alternativa), sólo que cuando la hipótesis nula es verdadera todos los τ_i son iguales a cero. Queda para el lector el desafío de realizar un esquema homólogo al de la Figura 1, pero que represente un caso en el que la hipótesis nula sea verdadera.

Surge de lo anterior que, las mismas hipótesis que acabamos de presentar, pueden ser planteadas como:

H_0: $\tau_i = 0$ para $i = 1,2,3$ equivalente a: ***H_0: $\mu_1 = \mu_2 = \mu_3$***

H_1: al menos un $\tau_i \neq 0$ equivalente a: ***H_1: al menos una μ_i es diferente***

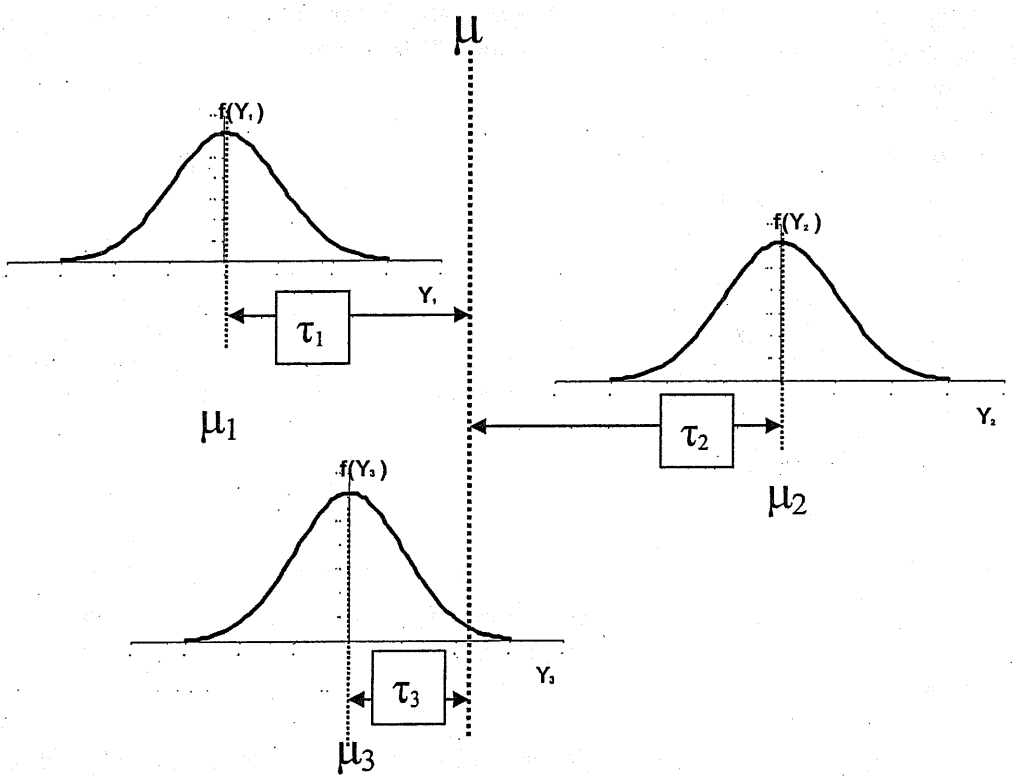
Figura 1. El modelo lineal aditivo y las hipótesis asociadas. El esquema representa las distribuciones poblacionales de la variable respuesta cuando la hipótesis nula es falsa, es decir cuando los distintos tratamientos producen efectos diferentes sobre la respuesta estudiada. Se recomienda al lector realizar un esquema semejante para representar una situación en que la hipótesis nula sea verdadera.

Respuesta = constante + efecto de tratamientos + error

$$Y_{ik} = \underbrace{\mu + \tau_i}_{\mu_i} + \varepsilon_{ik}$$

H_o: $\mu_1 = \mu_2 = \mu_3$ ➔ $\tau_i = 0$

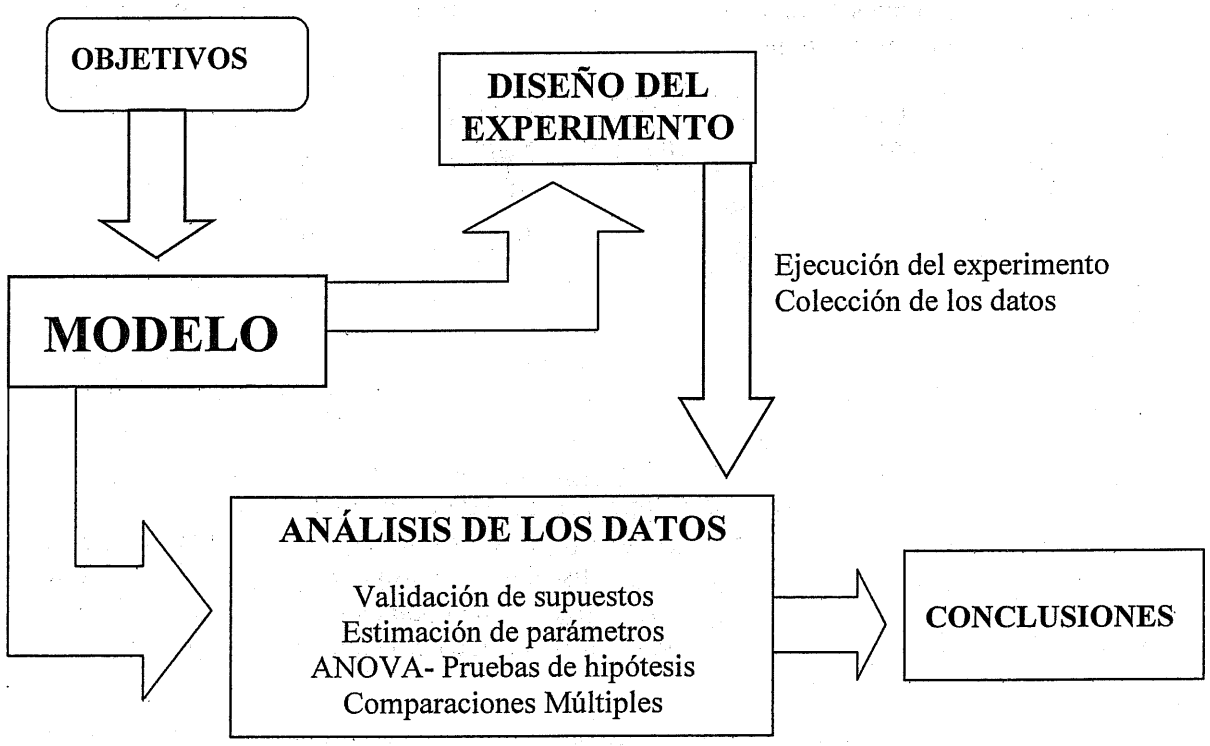
H_1: Al menos una μ_i es diferente



Estas hipótesis se ponen a prueba mediante el análisis de varianza de un solo factor que fue presentado en el capítulo anterior. Aparentemente nada cambió al desglosar la media de cada tratamiento en media general y efecto del tratamiento, ya que es absolutamente equivalente poner a prueba que las medias son iguales o que los efectos de los tratamientos son nulos. Pareciera que esta nueva formulación de las hipótesis asociada a los efectos de tratamientos que aparecen en el modelo es una complicación inútil ya que efectivamente en este caso no agrega ninguna información. La verdadera utilidad de separar los efectos asociados a distintas fuentes de variación recién se hace evidente cuando se tratan problemas más complejos como diseños en bloques para controlar la heterogeneidad del material experimental o diseños con tratamientos que combinan distintos factores. Sobre la misma estructura básica que acabamos de presentar, el modelo lineal aditivo se puede ampliar adicionando nuevos términos para incluir otras fuentes de variación planificadas en los experimentos.

En la formulación de cada modelo deben aparecer los parámetros poblacionales relevantes para el objetivo que impulsa el experimento, ya que el procedimiento que permite contestar las preguntas de interés es la prueba de hipótesis referida a los parámetros del modelo. De esta forma el modelo estadístico vincula los objetivos del estudio con el diseño del experimento y el análisis de los datos (Figura 2).

Figura 2: La complejidad del modelo y los parámetros a estimar son determinados por los objetivos del proyecto y al mismo tiempo son los determinantes de las fuentes de variación planificadas a ser consideradas en el diseño del experimento y en el análisis de los datos.



ESPERANZAS DE LOS CUADRADOS MEDIOS

Una vez que se ha planteado un modelo particular, es posible establecer qué parámetros se están estimando cuando se calculan los diferentes cuadrados medios y decidir en consecuencia cómo construir los cocientes de cuadrados medios que generen el valor de F adecuado para probar las hipótesis de interés. Para el modelo:

$$y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

donde se quiere poner a prueba la hipótesis que postula que los tratamientos no tienen efecto:

H_o: $\tau_i = 0$ para $i = 1,2,\ldots, t$

H_1: al menos un $\tau_i \neq 0$

Si el experimento se repitiera indefinidamente y se promediaran los cuadrados medios observados, obtendríamos los valores denominados en la tabla siguiente como ***valores esperados***. Para un experimento en particular se obtienen ***los cuadrados medios*** que ***son estimadores de esos valores esperados***. No vamos a presentar aquí la demostración matemática de esta afirmación, que se puede encontrar en numerosos textos, incluyendo la bibliografía sugerida al final del capítulo. Sin embargo, es importante subrayar que la demostración se fundamenta en los supuestos de independencia entre las observaciones y de homogeneidad de varianzas entre los tratamientos.

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	VALORES ESPERADOS
Tratamientos	t - 1	SC_{TRAT}	$SC_{TRAT}/ (t-1)$	$\sigma^2 + r.\Sigma \tau_i^2/ (t-1)$
Error	t (r – 1)	SC_E	$SC_E/ t (r-1)$	σ^2
Total	t r - 1	SC_{TOT}		

Cuando la hipótesis nula es verdadera, los cuadrados medios correspondientes a las dos fuentes de variación de este modelo estiman la misma varianza, ya que el término que contiene a la suma de los τ_i^2 es igual a cero. El fundamento del criterio de la prueba $F = CM_{TRAT}/CM_E$ es la comparación de dos varianzas muestrales independientes que son estimadores de la misma varianza común si la hipótesis nula es verdadera. En este punto se requiere asumir que las poblaciones siguen la distribución normal. Si se puede realizar este supuesto, la distribución del cociente de dos varianzas de poblaciones normales, cuando es verdadera la hipótesis nula (o sea que son poblaciones iguales) sigue una distribución de probabilidad F.

En cambio, cuando la hipótesis nula es falsa, algún τ_i es distinto de cero y el numerador de F se amplía dando lugar a un valor del cociente mayor al que se esperaría si la hipótesis

nula fuera verdadera. Valores grandes del cociente son poco usuales si la Ho es verdadera, con lo cual la posibilidad de detectar una diferencia aumenta a medida que la diferencia real entre los tratamientos se hace mayor. Esto sugiere que el cociente de cuadrados medios es un buen criterio para probar la hipótesis nula de que las medias poblacionales son las mismas en todos los tratamientos y que la región crítica estará situada en el extremo superior de la distribución. La denominada "distribución F" en realidad es una familia de distribuciones de probabilidad que difieren según los grados de libertad asociados al numerador y al denominador del cociente de cuadrados medios.

El ***nivel de significación α*** es, como en todas las pruebas de hipótesis, el umbral de probabilidad establecido para admitir que el resultado obtenido sea producto del azar, de manera que los tratamientos no tengan nada que ver con él. Esencialmente el argumento central es que mientras sea razonable explicar un fenómeno observado en términos de error experimental, no se sostiene que los tratamientos tienen efecto. Es habitual en biología utilizar α=0.05 como valor de probabilidad (P) al cual se quiebra el escepticismo y se concede que los tratamientos tienen efecto. Ciertamente no hay una gran diferencia entre un P=0.049 que conduce a decidir que los tratamientos tienen efecto y un P=0.051 que implicaría que no tienen ninguna eficacia. La tendencia actual de comunicar ***el valor P*** en lugar de sólo informar si es mayor o menor que α, parte de reconocer que existe un gradiente en la fuerza de la evidencia que los datos proveen para poner a prueba las hipótesis.

COMPARACIONES MÚLTIPLES

Cuando se rechaza la hipótesis nula de igualdad entre medias de tratamientos mediante el análisis de varianza, se desea conocer cuáles son las medias que difieren significativamente entre sí. En estos casos no podemos utilizar la prueba clásica *t* de comparación de pares de medias, para juzgar una por una su significación por separado. Esto es así porque el nivel de significación de las pruebas no permanece constante al realizar con los mismos datos comparaciones que no son independientes.

Existen varias pruebas de comparaciones múltiples que toman en cuenta el incremento de la probabilidad de error de tipo I en comparaciones repetidas y tienen como objetivo mantener en valores bajos la tasa de error del experimento completo, es decir la probabilidad de alcanzar al menos una falsa diferencia significativa. En general estos procedimientos funcionan eficientemente si previamente el factor en estudio resultó significativo en el ANOVA. Algunas de estas pruebas resultan demasiado conservadoras y en consecuencia presentan baja potencia, mientras otras son demasiado poco conservadoras. Si bien se han estudiado exhaustivamente sus propiedades asintóticas analíticamente y su comportamiento en conjuntos de datos simulados con propiedades conocidas, no se alcanzó aún consenso unánime en la comunidad científica acerca de cuál es "la mejor". En este curso presentaremos la prueba de comparaciones múltiples de Tukey porque es ampliamente utilizada en las ciencias agronómicas y porque sirve como ejemplo para la aplicación de cualquier otra de las pruebas similares. Para problemas de mayor complejidad que los presentados en este curso se recomienda revisar la bibliografía antes de seleccionar la prueba a posteriori más adecuada.

La prueba de Tukey examina con un mismo estadístico todas las diferencias de medias muestrales en estudio. Si hay t medias, luego habrá (t (t-1)/2) diferencias de medias posibles. El estadístico de Tukey es el siguiente:

$$DELTA = tuk_{(t,GLE,\alpha)}\sqrt{\frac{2CM_E}{r}}$$

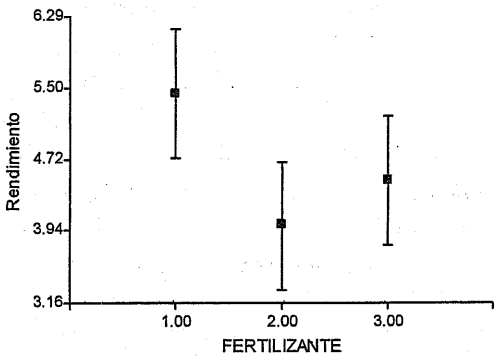
donde $tuk_{(t,GLE;\,\alpha)}$ es el cuantil (1-α) que se obtiene de la distribución ***de Tukey*** (ver Tabla en Anexo) para t tratamientos y los grados de libertad del error; **α** es el nivel de significación en base al cual se rechazó la H_0 del ANOVA y ***r*** es el número de repeticiones que contribuyen al cálculo de las medias muestrales. Si el valor absoluto de la diferencia entre un par de medias supera al valor Delta se dice que esta diferencia es estadísticamente significativa. Se concluirá en consecuencia que las medias poblacionales asociadas a esa diferencia son distintas con un nivel de significación α.

Retomando el problema de fertilización, recordemos que se había concluido que los diferentes fertilizantes diferían en efectividad. En la figura 3 se presenta una estimación por intervalo de confianza del 95% de las medias de los tres fertilizantes.

Aunque no es imprescindible, se puede construir una matriz de valores absolutos de las diferencias entre medias como la que se muestra a continuación.

TRATAMIENTO		*F2*	*F3*	*F1*
Promedio		**4**	**4.49**	**5.45**
Diferencias absolutas entre promedios	*F2*		0.49	1.45
	F3			0.96

Figura 3: Intervalos de confianza del 95% para los promedios del ejemplo de fertilización.



La pregunta que sigue es ¿cuál o cuáles de ellos producen rendimientos diferentes?. Para buscar una respuesta se utilizará el test de Tukey.

El segundo paso consiste en calcular el estadístico de Tukey. Para el ejemplo, t = 3, GL_E= 27 y alfa=0.05 (el mismo usado en el ANOVA), $tuk_{(3,\,27;\,\alpha}$ = 2.47; CM_E = 0.949 y r = 10 (número de repeticiones). Así se obtiene el estadístico Delta = 1.0803

Para completar el análisis basta controlar qué diferencias entre medias muestrales son mayores que 1.0803, para concluir entonces que las medias poblacionales a las que estiman difieren entre sí con un nivel de significación del 5%. Revisando la matriz de diferencias de medias se puede verificar que el fertilizante 1 difiere del 2 y ninguno de ellos difiere signi-

ficativamente del fertilizante 3. Este resultado se simboliza asignando letras distintas a los promedios que difieren significativamente, como se indica en la siguiente tabla.

TRATAMIENTO	*F2*	*F3*	*F1*
Promedio	**4**	**4.49**	**5.45**
Letras iguales corresponden a promedios que no difieren	**a**	**ab**	**b**

EL NÚMERO DE REPETICIONES

Por tradición en los experimentos agrícolas se han empleado tres o cuatro repeticiones sin considerar la importante influencia que tiene esta decisión sobre la capacidad del experimento para detectar diferencias significativas entre los tratamientos. Al incrementar el número de repeticiones aumenta siempre la precisión, aunque simultáneamente aumentan los costos de la experimentación, de manera que resulta de especial importancia encontrar el número conveniente de repeticiones para cada experimento.

La expresión matemática del valor Delta de Tukey permite estimar de manera aproximada el número de repeticiones adecuado para que el experimento sea capaz de detectar diferencias significativas de una deseada magnitud. Tal como se expresó más arriba, si el valor absoluto de la diferencia entre un par de medias supera al valor Delta se dice que esta diferencia es estadísticamente significativa. Al momento de diseñar un experimento, el investigador sabe si busca diferencias muy sutiles o más groseras y es capaz por lo tanto de aproximar un valor Delta deseado, de la misma manera que decide si la variable respuesta la medirá con un calibre o con un metro de carpintero. Si además cuenta con información de experimentos anteriores (propios o de trabajos publicados provenientes de sistemas semejantes) que le provea alguna estimación del CM_E (medida de la variabilidad de la respuesta) puede estimar r despejándolo de la expresión de Delta:

$$r = tuk^2 \cdot 2CM_E / DELTA^2$$

Queda claro que no se trata de un cálculo muy exacto, ya que la información necesaria para realizarlo no es conocida con precisión antes de realizar el experimento (por ejemplo la variabilidad de la variable respuesta). Sin embargo, es una aproximación que provee un buen punto de partida para comenzar a decidir el número apropiado de repeticiones en cada contexto y, sin duda, supera ampliamente a las decisiones arbitrarias de realizar unas pocas repeticiones para minimizar los costos de la investigación y encontrarse luego con los resultados de un experimento que sólo era capaz de detectar diferencias muy groseras entre tratamientos.

INTERVALOS DE CONFIANZA EN COMPARACIONES MÚLTIPLES

Se pueden construir intervalos de confianza para las diferencias entre medias en el contexto de las comparaciones múltiples. La forma del intervalo es igual a la de los intervalos para diferencias de medias poblacionales con muestras independientes, sólo que en este caso en lugar de la distribución *t* de Student se utiliza el valor de la distribución ***Tukey:***

$tuk_{(t,GLE;\alpha)}$ (el cuantil (1-α) que se obtiene para t tratamientos y los grados de libertad del error; donde 1-**α** es el nivel de confianza del intervalo) y también se utiliza la estimación de la varianza común a todos los tratamientos calculada en el ANOVA, o sea, CM_E, aquí nuevamente ***r*** es el número de repeticiones que contribuyen al cálculo de las medias muestrales.

$$IC_{i,j} \rightarrow \overline{y_i} - \overline{y_j} \pm tuk_{(\alpha,t,GL_E)} \sqrt{2\frac{CM_E}{r}}$$

Como siempre, es deseable que los intervalos resulten lo más estrechos posibles para un cierto nivel de confianza. Si observamos cómo se compone el término que se suma y se resta de la diferencia de medias muestrales, comprobamos que para un determinado nivel de significación, el intervalo será mas angosto si disminuye el CME (la variabilidad no controlada) y/o si aumenta el número de repeticiones.

$$P\left[IC_{i,j}cubra \rightarrow \mu_i - \mu_j\right] = 1-\alpha$$

En la tabla siguiente se presentan los intervalos de confianza simultáneos calculados para las diferencias de medias de rendimientos en el ejemplo de los fertilizantes

Diferencia de medias	Intervalo de Confianza del 95%
$\mu_1 - \mu_2$	(0.37, 2.53)
$\mu_1 - \mu_3$	(-0.12, 2.04)
$\mu_2 - \mu_3$	(-0.59, 1.57)

Se observa que los intervalos de confianza para las diferencias de medias que no resultaron significativas en las comparaciones múltiples comprenden al valor cero, como era esperable.

TASA DE DESCUBRIMIENTOS FALSOS

Existen estudios en los que se realiza un número muy grande de comparaciones simultáneas, como los que se conducen en investigaciones de genómica funcional, donde es necesario comparar el nivel de expresión de miles de genes (por ejemplo para seleccionar aquellos que mejor diferencian entre tejidos con distintos tipos de tumores). En estos casos los métodos de comparaciones múltiples que tienen como objetivo mantener la tasa de error del experimento completo (la probabilidad de alcanzar una o más falsas diferencias significativas), mayormente no permiten encontrar diferencia significativa alguna. Para reemplazarlos, en los últimos años se desarrollaron pruebas que controlan la proporción de hipótesis erróneamente rechazadas (tasa de descubrimientos falsos) en lugar de la probabilidad de alcanzar al menos una diferencia significativa espuria (Storey 2002, García 2003). Debido

a que estas técnicas presentan mayor potencia para series muy grandes de pruebas simultáneas, su uso se encuentra en gran expansión (Hopkins et al. 2002) y también el desarrollo de programas de computación para aplicarlas. Aunque todavía no fueron incorporadas como alternativas de análisis en los paquetes estadísticos comerciales, se encuentran programas para su implementación disponibles en la Web (ver en Storey and Tibshirani 2003).

Bibliografía citada

García, L.V. 2003. Controlling the false discovery rate in ecological research. Trends in Ecology and Evolution, 18: 553-554.

Hopkins, A.M. et al. 2002. A new source detection algorithm using the false discovery rate. Astronomy Journal, 123: 1086-1094.

Storey, J.D. 2002. A direct approach to false discovery rates. Journal of the Royal Statstical Society, 64: 499-517.

Storey, J.D. and Tibshirani, R. 2003. Statistical significance for genomewide studies. PNAS, 100 (16): 9440-9445.

Lecturas recomendadas

Kuehl, R. 2001. Diseño de experimentos. Segunda edición. Thomson. Capítulo 1, 2, 3.

Li, C.C. 1969. Introducción a la estadística experimental. Ediciones Omega, Barcelona. Capítulos 6, 7 y 8.

Steel, R. G. y J. H. Torrie. 1985. Bioestadística: Principios y Procedimientos. Segunda Edición. McGraw-Hill. Capítulo 7 y 8.

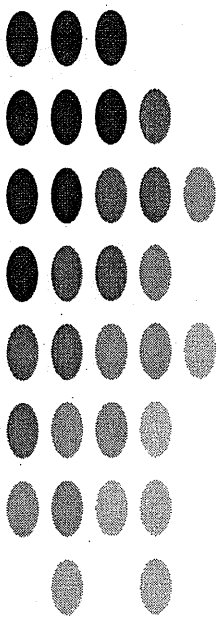
Underwood, A. J. 1997. Experiments in Ecology. Cambridge University Press, Cambridge. Capítulo 7.

Capítulo 4
El control de la heterogeneidad

Diferencias intrínsecas entre unidades pueden enmascarar el efecto de los tratamientos

Agrupar las unidades semejantes en bloques permite aislar en el análisis el efecto de la heterogeneidad

Diseños que reducen el error experimental



EL CONTROL DE LA HETEROGENEIDAD

En los dos capítulos anteriores se presentaron los elementos conceptuales fundamentales del análisis de varianza y de los modelos estadísticos asociados. En este capítulo se exponen dos diseños experimentales de amplia aplicación: el ***diseño completamente aleatorio*** y el ***diseño en bloques completos aleatorizados***. La diferencia entre estos diseños reside en la forma en que se asignan los tratamientos a las unidades experimentales, derivada del grado de heterogeneidad del material que se utiliza. En concordancia con el diseño, el modelo y el análisis de la varianza toman cuenta de estas diferencias.

DISEÑO COMPLETAMENTE ALEATORIO (DCA)

En el ***diseño completamente aleatorio*** se asume que las unidades experimentales son bastante homogéneas entre sí, es decir que no existe información disponible para clasificarlas en grupos que resulten más homogéneos que el conjunto total. El plan experimental consiste en asignar los tratamientos a las unidades enteramente al azar. El ensayo de fertilización que se discutió en los capítulos anteriores fue planeado según este diseño. En la Figura 1 se puede ver el plano de campo correspondiente a ese ensayo. Se trata de 30 parcelas con el mismo tipo de suelo, semejante nivel de fertilidad inicial y sembradas con el mismo cultivo en igual fecha y densidad de siembra. Se comprobó que la emergencia del cultivo resultó homogénea y al momento de fertilizar se numeraron las parcelas (izquierda arriba en el plano) y se les asignó a los primeros diez números al azar el fertilizante 1 (F1), a los siguientes diez el fertilizante 2 (F2) y a los restantes el fertilizante 3 (F3).

1 F3	2 F1	3 F3	4 F2	5 F1
6 F2	7 F3	8 F2	9 F1	10 F2
11 F3	12 F2	13 F1	14 F1	15 F3
16 F2	17 F3	18 F2	19 F2	20 F1
21 F1	22 F2	23 F1	24 F3	25 F2
26 F3	27 F1	28 F3	29 F1	30 F3

Figura 1: Plano de campo del experimento de fertilización con asignación al azar de los tratamientos. Los números a la izquierda arriba corresponden a la numeración de las parcelas, abajo se indica el fertilizante que se debe aplicar a cada una.

El análisis de los datos que resultan de un ***diseño completamente aleatorio*** es relativamente simple y la disminución de la información debida a la pérdida de datos en el transcurso de la experiencia (animales muertos, parcelas dañadas por ataque de enfermedades, material perdido accidentalmente) es menor que con cualquier otro plan experimental, ya que el número de repeticiones en cada tratamiento puede no ser igual (aunque es preferible que lo sea). Con respecto a otros diseños más complejos la principal desventaja que presenta el diseño completamente aleatorio es su falta de sensibilidad cuando el material utilizado en el ensayo no es uniforme. Esto se debe a que todas las fluctuaciones observadas entre las

respuestas de las distintas unidades se acumulan en el rubro "error" del análisis de la varianza. En consecuencia, si existe gran variación no controlada entre unidades, el ensayo sólo será capaz de detectar efectos muy groseros de los tratamientos. Por eso, cuando se busca detectar efectos más sutiles y existen fuentes conocidas *a priori* de variación en el material experimental, se recomiendan planes experimentales para la reducción del error experimental, como los diseños en bloques. En cambio, el diseño completamente aleatorio, se reserva para situaciones en las que la única fuente controlable de variación es la debida a los tratamientos.

Supongamos por ejemplo, que se conduce un experimento con 20 cerdos para comparar el aumento de peso logrado al alimentarlos con cuatro dietas distintas. Todos ellos provienen de una línea consanguínea y tienen la misma edad, peso, color, y sexo. Dado que no se pueden identificar entre estos animales grupos con características distintivas que pudieran determinar diferentes respuestas a los tratamientos, es razonable suponer que las diferencias observadas en ganancia de peso al finalizar el ensayo, podrán ser adjudicadas a las diferentes dietas. En consecuencia, se seleccionan al azar cinco cerdos para recibir cada tratamiento y se los alimenta individualmente. Cada cerdo constituye una unidad experimental y el ensayo tiene cinco repeticiones. En la Tabla 1 se presentan los resultados de este experimento.

Tabla 1: Aumentos de peso, expresados en kg /animal , de cerdos alimentados con 4 raciones diferentes

Ración	Repeticiones					Promedio
1	35	19	31	15	30	26
2	40	35	46	41	33	39
3	39	27	20	29	45	32
4	27	12	13	28	30	22

A continuación se presenta el modelo y el análisis de los datos de la Tabla 1. El modelo y el ANOVA corresponden exactamente al caso general de modelo lineal y análisis de la varianza para un solo factor de variación controlado por el experimentador, presentados en los capítulos anteriores. El lector puede saltear directamente al próximo parágrafo si no necesita una síntesis del tema.

El modelo es:

$$y_{ik} = \mu + \tau_i + \varepsilon_{ik}$$

los tratamientos	$i = 1, 2, 3, 4$	$t = 4$
las repeticiones	$k = 1, 2, 3, 4, 5$	$r = 5$

donde y_{ik} es el aumento correspondiente al **k**-ésimo cerdo que recibió la dieta **i** (por ejemplo $y_{32} = 27$ kg), μ es el aumento medio general, τ_i es el efecto de la dieta i, y ε_{ik} es el error correspondiente a la observación **k** de la ración **i**.

Supuestos del modelo: los ε_{ik} son variables aleatorias independientes $N(0; \sigma^2)$.

El residual e_{ik} se define como la desviación de una observación de la media del respectivo tratamiento:

$$e_{ik} = y_{ik} - \hat{y}_{ik} = y_{ik} - \bar{y}_{i.}$$

Por ejemplo el residual $\mathbf{e_{32}}$ = 27 kg – 32kg = -5kg

La hipótesis nula y la alternativa, que postulan que el aumento de peso no difiere entre animales que recibieron distintas raciones (H_o) o que al menos una ración produce un aumento de peso distinto a las demás (H_1), se expresan como:

H_o: $\tau_i = 0$ *para* $i = 1, 2, 3, 4$

H_1: *al menos un* $\tau_i \neq 0$

El nivel de significación de la prueba se establece como $\alpha = 0.05$
El análisis de varianza se presenta a continuación:

Cuadro de Análisis de la Varianza

F.V.	SC	gl	CM	F	valor P
Racion	823,75	3	274,58	3,99	0,0267
Error	1100,00	16	68,75		
Total	1923,75	19			

Como el valor P es igual a 0.0267, se concluye que existen diferencias significativas entre los aumentos de peso medios de los animales que recibieron las distintas raciones, es decir que al menos una ración difiere del resto.
El análisis continúa con el estudio de residuales para evaluar el cumplimiento de supuestos y una prueba de comparaciones múltiples para identificar cuál/es de las raciones son las que difieren.

DISEÑO EN BLOQUES COMPLETOS ALEATORIZADOS (DBCA)

Este diseño, de amplia aplicación en las ciencias agropecuarias, se fundamenta en el control de la heterogeneidad del material experimental mediante el agrupamiento de las unidades en bloques homogéneos.

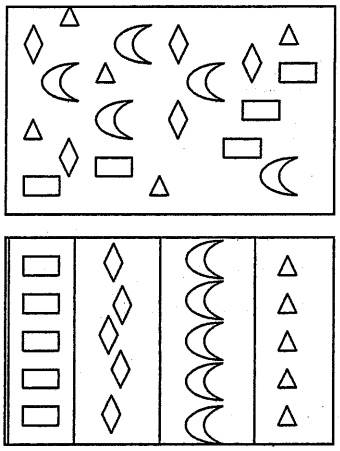


Figura 2: Se dispone de 20 unidades experimentales (animales, lotes, plantas, empresas o lagos, por ejemplo) para probar los efectos de cinco tratamientos. Estas unidades no son completamente iguales, son heterogéneas, tal como se simboliza mediante formas diferentes.

Agrupamos las unidades experimentales de acuerdo con sus características, de manera que todas las unidades iguales forman un grupo (bloque). En consecuencia quedan formados cuatro bloques, cada uno compuesto de 5 unidades homogéneas.

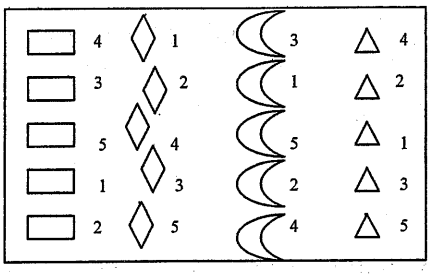


Figura 3: Dentro de cada bloque se asignan al azar los cinco tratamientos (indicados por los números del 1 al 5 en la figura). La evidente heterogeneidad entre las unidades está controlada por el proceso de formación de los bloques; la heterogeneidad oculta e incontrolable de las unidades se reparte al azar debido a que dentro de cada bloque los tratamientos se asignan al azar a las unidades.

En el ejemplo de la Figura 3, cada bloque tiene cinco unidades homogéneas y cada unidad del bloque recibe un tratamiento diferente. Los 5 tratamientos se repiten 4 veces, una vez en cada bloque. Es decir que cada bloque representa una *réplica completa del experimento*: una serie completa de los tratamientos que se prueban en el estudio. El principio de la formación de bloques se basa en que la variación entre los bloques resulte lo más grande posible y, por consiguiente, la variación dentro de los bloques sea mínima.

El diseño en bloques completos aleatorios permite separar en el análisis la variabilidad debida a los bloques, que se descuenta del rubro del error. De esta forma, si el bloqueo se realiza correctamente, el experimento resulta más sensible, ya que se reduce el CM_E ➔aumenta el cociente F de la prueba de hipótesis.

Modelo para el diseño en bloques completos aleatorios:

$$y_{ik} = \mu + \rho_k + \tau_i + \varepsilon_{ik}$$

i= 1, ..., t
k= 1, ..., r

donde

μ es la media general
ρ_k es el efecto del k-ésimo bloque (constante para cada bloque)
τ_i es el efecto del i-ésimo tratamiento (constante para cada tratamiento)
ε_{ik} es el error correspondiente a la observación del tratamiento i-ésimo del k-ésimo bloque

Supuestos del modelo:

1. Los ε_{ik} son variables aleatorias de error independientes $N(0; \sigma^2)$.

IMPORTANTE: Aquí aparece un nuevo supuesto:

2. Los efectos de bloque y tratamiento son aditivos: se supone los bloques no interactúan con la estructura de los tratamientos. Es decir que los bloques no tienen influencia sobre las diferencias entre los tratamientos; no hay interacción bloque-tratamiento.

La Tabla muestra la notación simbólica de los resultados muestrales de una investigación.

Nivel de Factor	Bloques					Totales Tratamientos	Medias Tratamientos
	1	...	k	...	r		
1	y_{11}	...	y_{1k}	...	y_{1r}	$y_{1.}$	$\bar{y}_{1.}$
.	.	...	.	...	.	.	
.	.	...	.	...	.	.	
i	y_{i1}	...	y_{ik}	...	y_{ir}	$y_{i.}$	$\bar{y}_{i.}$
.	.	...	.	...	.	.	.
.	.	...	.	...	.	.	.
t	y_{t1}	...	y_{tk}	...	y_{tr}	$y_{t.}$	$\bar{y}_{t.}$
Totales bloque	$y_{.1}$	...	$y_{.k}$	...	$y_{.r}$	$y_{..}$	
Medias bloque	$\bar{y}_{.1}$		$\bar{y}_{.2}$		$\bar{y}_{.r}$		$\bar{y}_{..}$

El total de los valores de las observaciones para el bloque k-ésimo es:

$$y_{.k} = \sum_{i=1}^{t} y_{ik}$$

La media del bloque k-ésimo es:

$$\bar{y}_{.k} = \frac{\sum_{i=1}^{t} y_{ik}}{t} = \frac{y_{.k}}{t}$$

Los parámetros del modelo y sus estimadores son:

Parámetro	**Estimador**
μ	$\bar{y}_{..}$
τ_i	$\hat{\tau}_i = \bar{y}_{i.} - \bar{y}_{..}$
ρ_k	$\hat{\rho}_k = \bar{y}_{.k} - \bar{y}_{..}$
σ^2	CM_E

En el DBCA el residual e_{ik} se define como:

$$e_{ik} = y_{ik} - \hat{y}_{ik} = y_{ik} - \bar{y}_{i.} - \bar{y}_{.k} + \bar{y}_{..}$$

Para explorar la aditividad Bloque-Tratamiento se pueden graficar los datos en función de los tratamientos, usando una partición por bloque e incluyendo todas las particiones en el mismo gráfico, y unir los puntos pertenecientes a un mismo bloque. Se obtendrán los perfiles por bloque. Si los perfiles se cruzan hay presencia de interacción.

Partición de la variabilidad total en el DBCA:

Como se ha mencionado, la medida convencional de la variabilidad total es la suma de cuadrados de las observaciones expresada como SC_{TOT} (suma de cuadrados totales)

$$SC_{TOT} = \sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{..})^2$$

Cuando utilizamos información acerca de los bloques, además de los tratamientos, la partición de la desviación total es la siguiente:

$$y_{ik} - \bar{y}_{..} = (\bar{y}_{.k} - \bar{y}_{..}) + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ik} - \bar{y}_{.k} - \bar{y}_{i.} + \bar{y}_{..})$$

Desviación Total	Desviación del bloque alrededor de la media general	Desviación del tratamiento alrededor de la media general	Desviación remanente

Así la desviación total puede ser vista, ahora, como la suma de tres componentes.

Cuando elevamos al cuadrado y luego realizamos las sumas para todas las observaciones:

$$\sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{..})^2 = \sum_{i=1}^{t}\sum_{k=1}^{r}\left[(\bar{y}_{.k} - \bar{y}_{..}) + (\bar{y}_{i.} - \bar{y}_{..}) + (y_{ik} - \bar{y}_{.k} - \bar{y}_{i.} + \bar{y}_{..})\right]^2$$

Desarrollando el cuadrado de la derecha, los términos de los productos cruzados se anulan por ser sumas de desviaciones con respecto a la media. Así:

$$\sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{..})^2 = t\sum_{k=1}^{r}(\bar{y}_{.k} - \bar{y}_{..}) + r\sum_{i=1}^{t}(\bar{y}_{i.} - \bar{y}_{..})^2 + \sum_{i=1}^{t}\sum_{k=1}^{r}(y_{ik} - \bar{y}_{.k} - \bar{y}_{i.} + \bar{y}_{..})^2$$

$$\Downarrow \qquad \Downarrow \qquad \Downarrow \qquad \Downarrow$$

$$SC_{Total} = SC_{Bloque} + SC_{Trat} + SC_E$$

Se puede demostrar que los grados de libertad asociados con las diferentes sumas de cuadrado son:

SC	GL
$SC_{BLOQUES}$	r-1
SC_{TRAT}	t-1
SC_E	(r-1)(t-1)
SC_{TOT}	tr - 1

A partir de las sumas de cuadrados, como ya vimos en el capítulo de Análisis de la Varianza, se obtienen los cuadrados medios.
De este modo, la tabla de ANOVA, incluyendo los cuadrados medios esperados para los efectos de tratamientos, es la siguiente:

FUENTE DE VARIACIÓN	GRADOS DE LIBERTAD	SUMA DE CUADRADOS	CUADRADOS MEDIOS	VALORES ESPERADOS
Bloques	r - 1	$SC_{BLOQUES}$	$SC_{BLOQUES}/(r-1)$	$\sigma^2 + t.\Sigma \rho_k^2/(r-1)$
Tratamientos	t - 1	SC_{TRAT}	$SC_{TRAT}/(t-1)$	$\sigma^2 + r.\Sigma \tau_i^2/(t-1)$
Error	(t-1) (r – 1)	SC_E	$SC_E/(t-1).(r-1)$	σ^2
Total	tr - 1	SC_{TOTAL}		

Así para probar la hipótesis acerca de los efectos de tratamientos:

$H_o: \tau_i = 0$ *para* $i = 1,2,\ldots, t$

$H_1:$ *al menos un* $\tau_i \neq 0$

Utiliza como estadístico de prueba:

$$F = \frac{CM_{TRAT}}{CM_E}$$

Que se distribuye, bajo la hipótesis nula, como una F con t-1 y (t-1)(r-1) grados de libertad.

Ejemplo:

Se realizó una experiencia para estudiar el efecto de dosis crecientes de vitamina B12 en la ración alimentaria que se da a los cerdos. Como no se disponía de animales con características homogéneas en cantidad suficiente se aplicaron a distintas camadas de cerdos similares con respecto a edad, sexo y peso inicial. Los resultados obtenidos, que expresan el aumento de peso en kg son:

Tratamientos	**Camadas**			
	I	**II**	**III**	**IV**
Dieta 1	15	12	24	22
Dieta 2	22	16	26	24
Dieta 3	33	28	44	37
Dieta 4	31	24	40	34
Dieta 5	25	22	33	27

$H_o: \tau_i = 0$ *para* $i = 1,2,\ldots, 5$

$H_1:$ *al menos un* $\tau_i \neq 0$

Nivel de significación, α= 0,05

Análisis de la varianza

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	valor p
BLOQUE	454,95	3	151,65	56,34	<0,0001
DIETA	805,70	4	201,43	74,83	<0,0001
Error	32,30	12	2,69		
Total	1292,95	19			

Como el valor p< 0,0001 es menor que el nivel de significación prefijado, α= 0,05, se rechaza la hipótesis de que todas las dietas producen el mismo aumento promedio de peso.

El análisis continúa con el estudio de los residuales para verificar el cumplimiento de los supuestos y con la prueba de Tukey para discernir cuál/es dieta/s recomendar.

EFECTOS CONFUNDIDOS

Hay buenos y malos diseños. A veces un mal diseño obliga a un análisis más complicado de los datos, pero si es demasiado malo impide completamente extraer información interesante del estudio, esto ocurre cuando hay efectos confundidos. Como ejemplo extremo imaginemos un experimento en el que se alimenta a 200 niños japoneses con pescado y a 200 niños suecos, de la misma edad, con carne durante un mismo período de tiempo, supongamos cuatro años. Los resultados son: altura media de los niños japoneses 1.55m y altura media de los niños suecos 1.68m. Un nutricionista concluye que las diferencias de altura se deben a la dieta, un antropólogo objeta que posiblemente la mayor parte de la diferencia refleje diferencias étnicas. Cada uno de ellos puede tener razón o estar equivocado. No hay forma de saberlo porque los efectos están *confundidos* y no existe el análisis que permita separarlos.

EL CONTROL DE LA HETEROGENEIDAD

El rubro "error" abarca todas las causas de variación que no se controlaron durante el ensayo. Entre éstas algunas no se controlaron porque son incontrolables por naturaleza. Todo material biológico, por homogéneo que sea, presenta una cierta fluctuación cuyos factores determinantes no se conocen y son por lo tanto incontrolables y además las observaciones están sujetas a errores de medición, que deben ser minimizados dentro de lo posible. En cambio, otras fuentes de variabilidad podrían ser controladas. Por ejemplo en experimentos con animales se conoce que ciertos caracteres cuantitativos difieren entre el macho y la hembra, o entre animales de distintas camadas o de distinta edad o de diferente raza. En un diseño completamente al azar que utiliza animales de distintas edades, camadas, razas y sexos, toda esa variación contribuye al rubro del error. Una alternativa para reducir el error experimental es aplicar los tratamientos sobre animales *tan homogéneos como sea posible* (lo mismo se aplica a parcelas, plantas, empresas o cualquier otro tipo de unidades experimentales). En esos casos se elimina la heterogeneidad *antes*, y se aplica un diseño completamente aleatorio, pero se debe tener en cuenta que las conclusiones de tales ensayos sólo serán aplicables a condiciones tan restringidas como las que se establecieron al

seleccionar el material experimental. Es decir que *al acotar la variación del material que participa en el ensayo se restringe la población de referencia a la que apuntan las hipótesis*. Por ejemplo, se puede concluir que una ración produce el mayor aumento de peso en cerdos de determinada raza, edad, sexo, etc.

Uno de los descubrimientos más originales de la estadística es que se puede eliminar del error un factor controlable de variación *después* de haber realizado el ensayo usando material no homogéneo, mediante experimentos con aleatorización restringida, como los diseños en bloques. Los factores de variación quedan controlados porque se concibe un plan de experimentación de tal forma que resulte posible su eliminación ulterior en el cálculo del análisis de la varianza. Estos diseños permiten poner a prueba hipótesis referidas a poblaciones más heterogéneas y, en consecuencia, más amplias, sin provocar aumentos del cuadrado medio del error que resten sensibilidad al experimento.

Lecturas recomendadas

Kuehl, R. 2001. Diseño de experimentos. Segunda edición. Editorial Thomson.

Li, C.C. 1969. Introducción a la estadística experimental. Ediciones Omega, Barcelona.

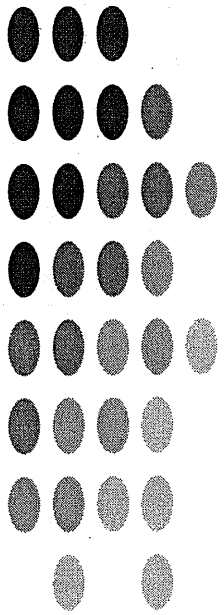
Steel, R.G. y J.H. Torrie. 1985. Bioestadística: Principios y Procedimientos. Segunda Edición. McGraw-Hill.

Underwood, A.J. 1997. Experiments in Ecology. Cambridge University Press, Cambridge.

Capítulo 5
Validación de supuestos

Supuestos que deben cumplirse para que los resultados sean confiables

Procedimientos para diagnosticar la violación de supuestos



VALIDACIÓN DE SUPUESTOS DEL ANOVA

Para que podamos confiar en los resultados de los análisis que realizamos, necesitamos verificar si los supuestos sobre los que se fundamenta el procedimiento de análisis se cumplen en el conjunto de datos que estamos estudiando. En orden decreciente de importancia, los supuestos del ANOVA que dan validez a la interpretación de los resultados son:

1. **Independencia de los errores experimentales**. Para garantizar el cumplimiento de este supuesto se deben asignar mediante algún procedimiento de azar los tratamientos a las unidades experimentales y se debe cuidar que las repeticiones de los distintos tratamientos permanezcan independientes durante la ejecución del experimento.

2. **Varianzas homogéneas** entre tratamientos. Es decir que las subpoblaciones bajo estudio, aunque difieran en su media, tienen la misma varianza, cuyo estimador es el CM_E.

3. **Distribución normal** de la variable respuesta. La razón por la cual suponemos normalidad, se debe a que de esta manera conocemos la distribución de probabilidad del cociente F y podemos utilizarla para cuantificar la incertidumbre de nuestras conclusiones.

Existen procedimientos para poner a prueba la violación de estos supuestos y alternativas de análisis de los datos para algunos de los casos en los que no se cumplen. Si bien un tratamiento completo de este tema excede el nivel de nuestro curso, presentaremos a continuación algunos de sus aspectos más sencillos. El orden decreciente de importancia o gravedad establecido para la violación de estos supuestos, se refiere a la disponibilidad de herramientas alternativas para el análisis de los datos cuando no se cumplen y también al impacto que tienen sobre las pruebas de hipótesis asociadas al ANOVA.

El análisis de la varianza se puede realizar con tamaños muestrales iguales o distintos, sin embargo, es recomendable utilizar iguales tamaños por dos motivos:

1) La prueba *F* es insensible a moderadas violaciones del supuesto de varianzas iguales entre tratamientos si el tamaño de las muestras es igual.

2) Igual tamaño de muestras minimiza la probabilidad de error de tipo II.

En este curso trataremos sólo el caso de los diseños con igual número de repeticiones, que se denominan ***diseños balanceados***.

Los métodos que permiten diagnosticar la violación de supuestos en un conjunto de datos comprenden tanto a pruebas estadísticas formales así como a procedimientos informales que consisten en explorar gráficamente los residuales. Recordemos que los residuales se obtienen por diferencia entre los valores observados de la variable respuesta y los respectivos valores predichos por el modelo. Es decir que los residuos constituyen una nueva variable aleatoria que puede ser usada para generar medidas y/o gráficos de diagnóstico para la verificación de supuestos y/o pruebas de adecuación del modelo.

Dado que el error experimental es la única porción aleatoria de la variable respuesta, la aplicación válida de las pruebas de hipótesis en el ANOVA implica que los errores experimentales sean independientes, normalmente distribuidos y con varianzas homogéneas para todas las observaciones.

A continuación se presentan algunas estrategias que pueden conducirse en InfoStat para verificar los supuestos.

Normalidad: Para analizar la normalidad de los residuales, del modelo de ANOVA propuesto, se puede construir un gráfico Q-Q plot normal tomando como variable a los residuales obtenidos en el análisis. Este gráfico muestra un diagrama de dispersión de los residuales en función de los cuantiles teóricos de una distribución normal estándar. Si los puntos se alinean sobre una recta de 45°, el gráfico sugiere que la distribución de los residuos es normal. Acompaña al gráfico el estadístico de Shapiro. Esta medida corresponde al cuadrado del coeficiente de correlación obtenido entre los residuos y los estadísticos de orden muestrales. Valores cercanos a 1 sugieren distribución normal para la variable en estudio.

Homogeneidad de varianzas: Para verificar la homogeneidad de varianzas puede realizarse un gráfico de dispersión de los residuales en función de los valores predichos por el modelo. Si la nube de puntos no presenta patrón alguno, se puede decir que hay homogeneidad. Los patrones que muestran forma de cono (a medida que crecen los valores predichos, los residuales son mayores), son casos típicos de falta de homogeneidad.

La homogeneidad de varianzas se puede probar haciendo un análisis de la varianza usando como variable dependiente al valor absoluto de los residuales (prueba de Levene). InfoStat permite guardar automáticamente dichos valores absolutos en la Solapa Modelo de la ventana del Análisis de la Varianza.

Independencia: Para verificar el supuesto de independencia (los errores son independientes o con covarianza cero), se puede realizar un gráfico de dispersión de los residuales en función de la variable que se piensa puede ocasionar dependencia entre las observaciones como, por ejemplo, la secuencia en el tiempo en que se realizaron las observaciones. La estructura de dependencia puede estar asociada a la forma en que se recolectaron los datos. Una tendencia a tener agrupados residuales positivos y/o negativos indica la presencia de correlación o falta de independencia. En general, un proceso de buena **aleatorización** asegura la independencia.

En determinados estudios las observaciones están inevitablemente correlacionadas. Esto ocurre por ejemplo cuando se evalúan en distintas fechas características de las mismas plantas porque el objetivo del trabajo justamente apunta a conocer la evolución de esa característica en el tiempo en los mismos organismos. Lo mismo ocurre cuando se estudian tendencias en los precios o en las precipitaciones a través del tiempo, o cuando se mide la cobertura de una maleza en cuadrados contiguos para evaluar su dispersión en el espacio. Para estos casos se formulan modelos más complejos que contemplan la covariación (variación conjunta debida a la falta de independencia) entre las observaciones y permiten estimarla. Esos modelos no serán abordados en este curso.

Fuera de estos casos particulares, en los cuales estimar la estructura de correlación es un objetivo del trabajo, el problema de falta de independencia entre las observaciones es el más difícil de salvar, de allí la gran importancia que se le asigna a la aleatorización en el diseño de los experimentos. Los problemas de falta de normalidad y de heterogeneidad de varianzas, en cambio, tienen distintas alternativas de solución que abarcan desde cambiar la escala de medición de la variable respuesta a través de transformaciones, hasta la aplicación de otros métodos de análisis, como es el caso de los métodos no paramétricos que no exigen el supuesto de normalidad. Las transformaciones más usuales son la raíz cuadrada, la logarítmica y la transformación angular. Una vez transformados los datos se verifica el cumplimiento de los supuestos sobre las observaciones en la escala transformada y si los problemas se solucionaron se procede al análisis en esa escala. Finalmente se retransforman los estimadores para presentarlos en su escala original. A continuación se presenta un ejemplo siguiendo el protocolo completo.

Ejemplo 1: Supuestos de Homogeneidad de Varianzas y Normalidad

Se tomaron 11 muestras aleatorias en el centro de cada uno de tres lagos, a una profundidad de 50 cm. Se determinó el contenido medio de oxigeno disuelto (en ppm) y se obtuvieron los datos que se presentan en la siguiente tabla. Con esta información se espera poder contestar si la concentración media de oxígeno difiere entre los tres lagos:

MUESTRA	*LAGO 1*	*LAGO 2*	*LAGO 3*
1	1	2	3
2	0	1	14
3	2	3	26
4	1	4	25
5	2	6	18
6	1	8	19
7	2	7	22
8	3	5	21
9	4	3	16
10	1	4	20
11	5	5	30

Los valores medios y de dispersión, estimados para cada lago son:

LAGO	*MEDIA Oxígeno*	*r*	*DESVIO ESTANDAR*
1	2.0000	11	1.4832
2	4.3636	11	2.1106
3	19.455	11	7.1325

Con esta información se realizó un análisis de comparación de medias de contenido de oxígeno disuelto en el agua (variable original) entre los tres lagos con un modelo lineal simple que asume un diseño completamente aleatorio (DCA), aquí no se muestran los resultados de ese análisis. Se guardaron los residuales en valor absoluto (que denominaremos variable Z) de ese análisis, para verificar el cumplimiento de los supuestos antes de interpretar los resultados de el análisis de comparación de promedios. Aquí sólo nos concentraremos en la validación de supuestos a través del análisis de residuales.

Con los residuales (variable Z) se realizo el análisis de Levene para estudiar la homogeneidad de varianzas entre tratamientos. La tabla de ANOVA obtenida con la prueba de Levene se presenta a continuación:

ANALISIS DE VARIANZA PARA Z (Prueba de Levene para residuales del análisis con los datos originales)

FUENTE DE VARIACION	*GL*	*SUMA DE CUADRADOS*	*CUADRADOS MEDIOS*	*F*	*VALOR P*
LAGO	2	95.7496	47.8748	5.50	0.0092
RESIDUAL	30	261.052	8.70173		
TOTAL	32	356.801			

Dado que se rechazó el supuesto de homogeneidad de varianzas se propone utilizar para el análisis de comparación de promedios a la variable transformada: raíz cuadrada de las observaciones = $Y_i^{1/2}$. En las figuras de la página siguiente se muestran los histogramas de frecuencias para la variable original y la transformada, allí se observa que la última presenta una distribución menos asimétrica.

En la tabla siguiente se presentan los valores medios y de dispersión para cada lago utilizando la variable transformada (raíz cuadrada de las observaciones originales = $Y_i^{1/2}$):

LAGO	*MEDIA*	*r*	*DESVIO ESTANDAR*
1	1.2919	11	0.6034
2	2.0249	11	0.5382
3	4.3088	11	0.9888

Una vez realizado el ANOVA de comparación de promedios sobre la variable transformada ($Y_i^{1/2}$), se conservan nuevamente los residuales del análisis (que denominaremos en este caso variable T) para poner a prueba la homogeneidad de varianzas en la variable transformada.

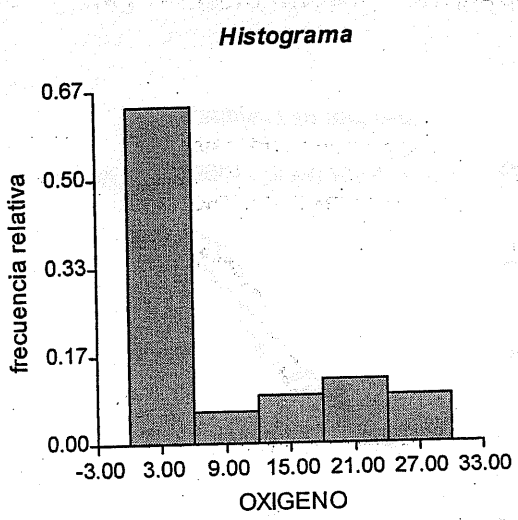
ANALISIS DE VARIANZA PARA T (Prueba de Levene para residuales del análisis con los datos transformados)

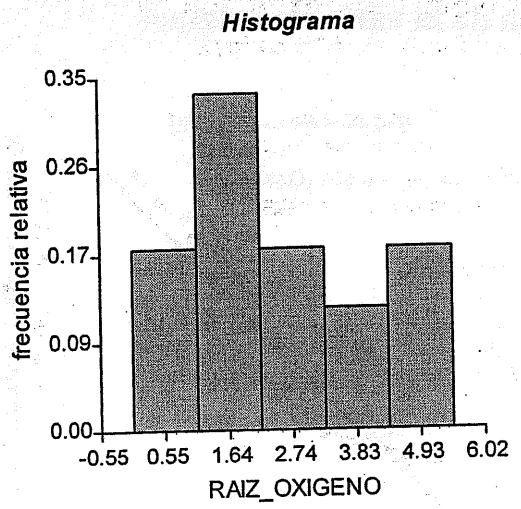
FUENTE DE VARIACION	*GL*	*SUMA DE CUADRADOS*	*CUADRADOS MEDIOS*	*F*	*VALOR-P*
LAGO	2	0.32898	0.16449	0.64	0.5355
RESIDUAL	30	7.73837	0.25795		
TOTAL	32	8.06736			

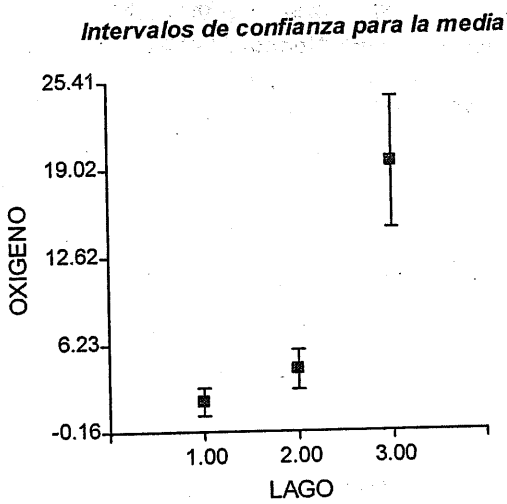
No se rechaza Ho. Es decir que en el caso de la variable transformada no hay evidencias de heterogeneidad de varianzas entre tratamientos. Esta secuencia de pruebas estadísticas nos permite verificar que la variable más apropiada (en el sentido del cumplimiento de los supuestos del ANOVA) para comparar los promedios de concentración de oxígeno en estos lagos, es la variable transformada ($Y_i^{1/2}$). Una vez que se ha comprobado que esta variable cumple con los supuestos de homogeneidad de varianzas y normalidad, la pregunta principal del estudio acerca de los promedios de concentración de oxígeno en los lagos se contesta con la prueba de hipótesis de comparación de medias de la variable transformada.

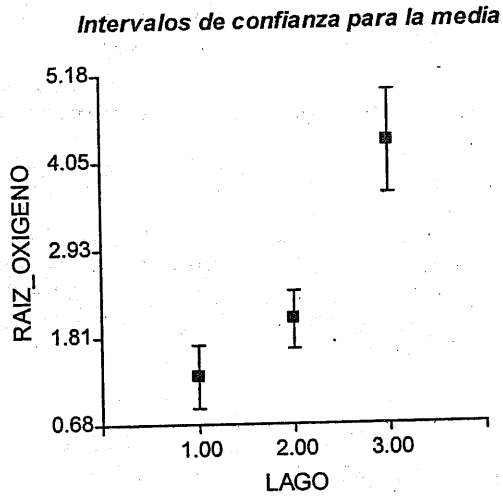
Variable original: oxígeno en ppm

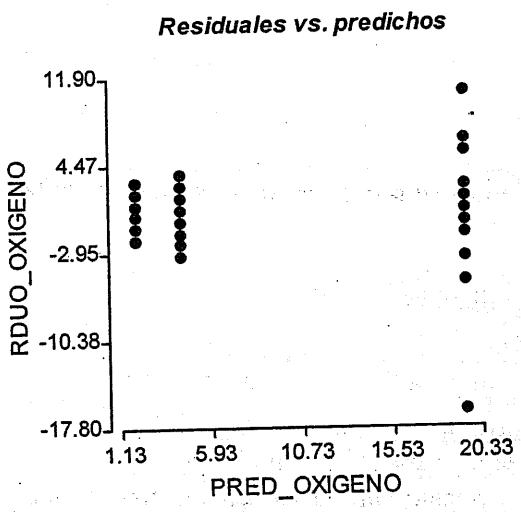
Variable transformada = raíz cuadrada de la variable original

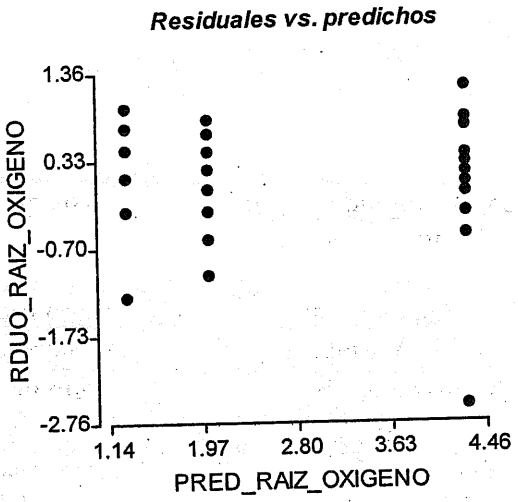






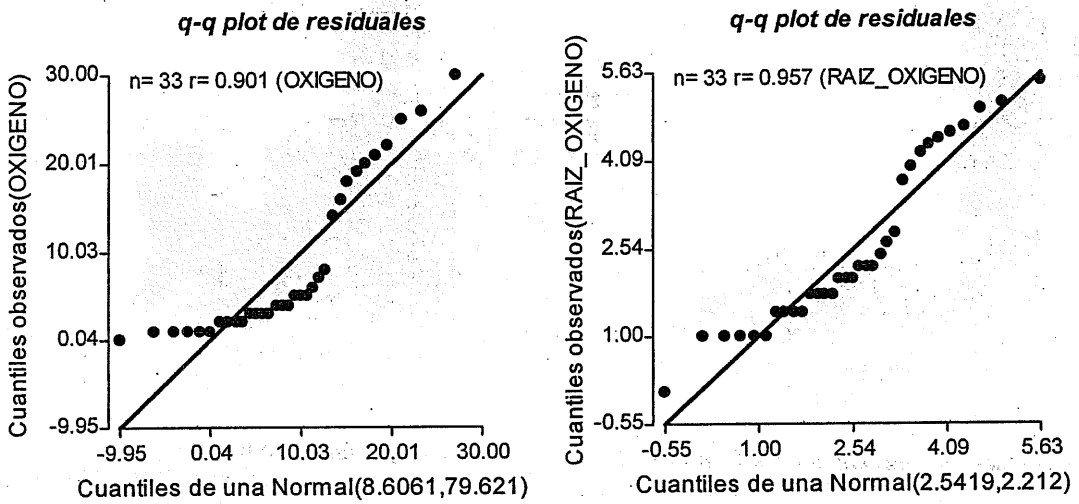






El siguiente q-q plot muestra que los datos transformados también presentan distribución más cercana a la normal.

Variable original: oxígeno en ppm **Variable transformada = raíz cuadrada de la variable original**



Análisis de la varianza para la variable transformada

Variable	N	R^2	R^2 Aj	CV
RAIZ_OXIGENO	33	0.77	0.75	29.01

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	54.47	2	27.23	50.07	<0.0001
LAGO	54.47	2	27.23	50.07	<0.0001
Error	16.32	30	0.54		
Total	70.78	32			

Test:Tukey Alfa:=0.05 DMS:=0.77514

Error: 0.5439 gl: 30

LAGO	Medias	n		
1.00	1.29	11	A	
2.00	2.02	11	A	
3.00	4.31	11		B

Letras distintas indican diferencias significativas(p<= 0.05)

Los datos utilizados en todos los análisis precedentes se presentan en la siguiente tabla, donde las variables son:

LAGO: Variable Categórica que representa a los distintos lagos
OXIGENO: Variable observada expresada en ppm (Yi)
PRED: Predicho obtenido con el modelo propuesto para la variable original ($\hat{Y}_i$)
RESIDUAL: Residual obtenido con el modelo propuesto para la variable original (Z_i)
RAIZOXI: .Variable transformada (raíz cuadrada de variable original = $Y_i^{1/2}$)
RAIZPRED Predicho obtenido con el modelo propuesto para la variable transformada
RAIZRES: Residual obtenido con el modelo propuesto para la variable transformada

LAGO	OXIGENO	PRED	RESIDUAL	RAIZOXI	RAIZPRED	RAIZRES
1	1	2	-1	1	1.292	-0.291
1	0	2	-2	0	1.292	-1.291
1	2	2	0	1.414	1.292	0.122
1	1	2	-1	1	1.292	-0.291
1	2	2	0	1.414	1.292	0.122
1	1	2	-1	1	1.292	-0.291
1	2	2	0	1.414	1.292	0.122
1	3	2	1	1.732	1.292	0.440
1	4	2	2	2	1.292	0.708
1	1	2	-1	1	1.292	-0.291
1	5	2	3	2.236	1.292	0.944
2	2	4.364	-2.363	1.414	2.025	-0.610
2	1	4.364	-3.363	1	2.025	-1.024
2	3	4.364	-1.363	1.732	2.025	-0.292
2	4	4.364	-0.363	2	2.025	-0.024
2	6	4.364	1.6363	2.449	2.025	0.424
2	8	4.364	3.6363	2.828	2.025	0.803
2	7	4.364	2.6363	2.645	2.025	0.620
2	5	4.364	0.6363	2.236	2.025	0.211
2	3	4.364	-1.363	1.732	2.025	-0.292
2	4	4.364	-0.363	2	2.025	-0.024
2	5	4.364	0.636	2.236	2.025	0.211
3	3	19.455	-16.454	1.732	4.309	-2.576
3	14	19.455	-5.454	3.741	4.309	-0.567
3	26	19.455	6.545	5.099	4.309	0.790
3	25	19.455	5.545	5	4.309	0.691
3	18	19.455	-1.454	4.242	4.309	-0.066
3	19	19.455	-0.454	4.358	4.309	0.050
3	22	19.455	2.545	4.690	4.309	0.381
3	21	19.455	1.545	4.582	4.309	0.273
3	16	19.455	-3.454	4	4.309	-0.308
3	20	19.455	0.545	4.472	4.309	0.163
3	30	19.455	10.545	5.477	4.309	1.168

Ejemplo 2: Supuesto de independencia entre bloques y tratamientos

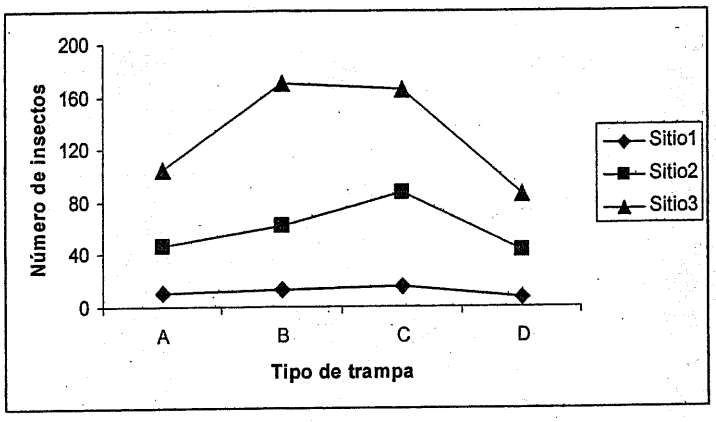
En los diseños que utilizan bloques para controlar la heterogeneidad del material experimental, se debe tener en cuenta un supuesto más para el análisis de varianza: la ausencia de interacciones bloque-tratamiento.

La presencia de interacciones bloque–tratamiento puede ser detectada a partir de un gráfico de las respuestas por bloques. Una severa falta de paralelismo en el gráfico indica que los bloques y los tratamientos interactúan en sus efectos en la variable respuesta. Veamos el siguiente ejemplo:

En un estudio de monitoreo de insectos se compararon 4 tipos distintos de trampas que se colocaron aleatoriamente en cada uno de tres sitios (bloques) que presentaban distinta cantidad de insectos. El diferente nivel de infestación fue la variable de bloqueo para el diseño. Los resultados en número de insectos fueron los siguientes:

Tipo de trampa	Sitio1	Sitio2	Sitio3
A	11	47	105
B	13	62	170
C	15	87	165
D	7	43	86

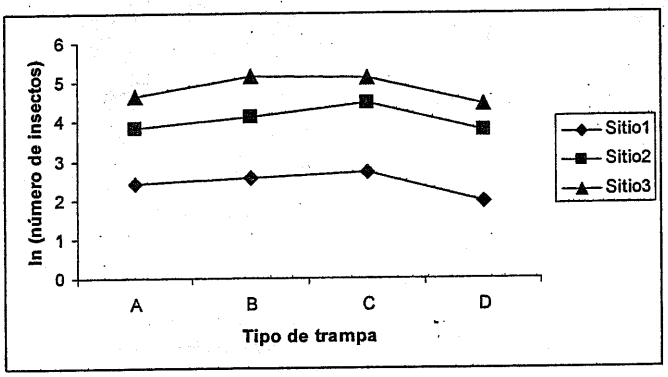
Para verificar la ausencia de interacciones bloque-tratamiento graficamos la variable respuesta (número de insectos atrapados) para los distintos tratamientos (trampas) en cada bloque.



En el gráfico podemos ver falta de paralelismo de las líneas, lo que estaría indicando presencia de interacciones.

Una posible solución a este problema es la transformación de los datos mediante el cálculo de logaritmos. Esta transformación tiene la particularidad de que al aplicar logaritmos a un producto lo transforma en suma de logaritmos. De este modo los efectos multiplicativos (interacción) se transforman en aditivos.

En el siguiente gráfico se muestran las mismas relaciones para los datos transformados ($\ln Y_i$), allí se puede ver que las líneas son casi paralelas lo que da indicios de no interacción entre los efectos de sitios y trampas.



El análisis de varianza para los datos transformados que se presenta a continuación, permite rechazar la hipótesis de igualdad de medias. Se concluye que existen diferencias significativas entre trampas en el logaritmo del número de insectos capturados.

Cuadro de Análisis de la Varianza para ln del número de insectos

F.V.	SC	gl	CM	F	p-valor
sitio	12,34	2	6,17	498,33	<0,0001
trampa	0,90	3	0,30	24,18	0,0009
Error	0,07	6	0,01		
Total	13,31	11			

En los dos ejemplos presentados, un cambio de escala en la variable respuesta estudiada, mediante la transformación raíz cuadrada en el primer caso y logaritmo natural en el segundo, resultó en una nueva variable que cumple los supuestos del ANOVA. Una vez que se verifica el cumplimiento de supuestos en la variable transformada, lo cual nos asegura que el riesgo de cometer un error de tipo I será igual al valor alfa que fijemos para las pruebas, tanto el análisis posterior como las conclusiones se refieren a la variable transformada.

Capítulo 6
Los experimentos factoriales

Experimentos que permiten responder preguntas acerca de sistemas complejos.

Estudio de factores que actúan simultáneamente y que pueden interactuar

de gran relevancia práctica, porque si la respuesta observada es consistente a través de las distintas condiciones de suelo, la amplitud de las conclusiones será mucho mayor. En cambio, si no hay un efecto uniforme del fungicida a través de los distintos suelos, interesará preguntar...

c) ¿Cuál es la mejor combinación de niveles de fungicida (con o sin) y de tipo de suelo, para la emergencia de trébol blanco?

Supongamos que se realiza el mismo experimento también con alfalfa, que es otra especie leguminosa utilizada como forrajera y, antes de presentar el modelo y el análisis formal de los experimentos factoriales, detengámonos un momento a comparar los resultados de ambos experimentos presentados como promedios de tres repeticiones para cada tratamiento en las figuras 1 y 2.

Figura 1: emergencia promedio de plantas de trébol en distintos tipos de suelo combinados con tratamientos de fungicida (con y sin aplicación).

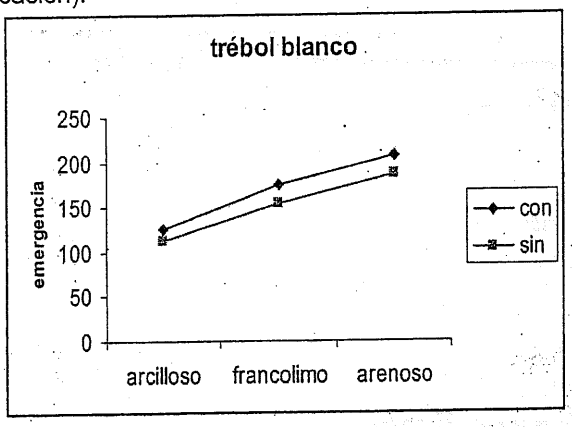
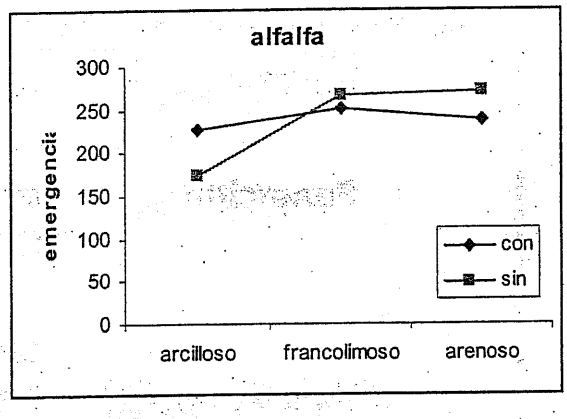


Figura 2: emergencia promedio de plantas de alfalfa en distintos tipos de suelo combinados con tratamientos de fungicida (con y sin aplicación).



En trébol blanco se observa que en promedio la emergencia es mayor en las macetas tratadas con fungicida en todos los suelos y también se observa que la emergencia es mayor en los suelos arenosos que en los suelos francos y en estos es mayor que en los arcillosos, tendencia que no difiere si los suelos fueron tratados o no con fungicida. Aparentemente, el efecto del fungicida no cambia entre los distintos suelos y el efecto de los suelos tampoco varía entre los distintos niveles del factor fungicida. Cuando la respuesta a un factor es la misma en los distintos niveles del otro, se dice que **los factores actúan de forma independiente** o que no interactúan. Si los factores no interactúan, nos interesa estudiar el efecto de cada factor promediado a través de los distintos niveles del otro, lo que se denomina **efecto principal** del factor. En este caso por ejemplo, un estimador del efecto principal del fungicida es la diferencia observada en la emergencia de plantas entre macetas tratadas con fungicida , promediada a través de los distintos suelos con respecto a la respuesta promedio de emergencia observada en todo el experimento.

Tabla 1: Promedio de emergencia de plantas de trébol blanco en cada tratamiento (combinación fungicida-suelo) a partir de 300 semillas sembradas en cada maceta.

			FACTOR A (fungicida)		
			a1: con	a2: sin	media
FACTOR B (Tipo de suelo)	arcilloso	b1	126.67	112	119.33
	franco	b2	175.67	154	164.83
	arenoso	b3	206.67	187	196.83
	media		169.67	151	160.33

Efecto principal de a_1= 169.67 - 160.33 = 9.34

Es decir que el agregado de fungicida produce un aumento en la emergencia promedio de aproximadamente 9 plantas. Evidentemente sólo tiene sentido evaluar el efecto promedio del fungicida combinando los resultados de los diferentes suelos, cuando podemos suponer que el fungicida no actúa de manera diferente en cada suelo, es decir cuando **no hay interacción** significativa entre los factores fungicida y suelo.

En cambio, en los resultados del experimento de alfalfa (Figura 2 y Tabla 2) se observa que el fungicida tiene mayor efecto sobre la emergencia en suelos arcillosos que en los otros tipos de suelo, incluso en suelos arenosos se observa una menor emergencia promedio en las macetas con fungicida que en las que no se aplicó fungicida, es decir que el efecto del fungicida varía según el tipo de suelo y al mismo tiempo el efecto del tipo de suelo es diferente si fueron tratados o no con fungicida (observar que en los tratados hay menor efecto del tipo de suelo sobre la emergencia que en los no tratados). Si se confirma que esta tendencia es significativa (luego presentaremos el análisis formal que nos permitirá poner a prueba las hipótesis de interés) se dice que los factores no actúan en forma independiente, es decir que existe ***interacción*** entre los factores. Una vez detectada una interacción significativa, carece de sentido el estudio de los efectos principales de los factores de manera independiente, ya que al promediar entre niveles donde la respuesta es diferente se enmascaran los efectos simples que tienen lugar dentro de cada nivel del otro factor.

Si calculamos en este caso el efecto principal de a_1 veremos que produce una información muy confusa ya que promedia el efecto del fungicida en los tres tipos de suelo y para esta especie el efecto del fungicida es diferente en los distintos suelos (aumenta la emergencia en suelos arcillosos y la disminuye en arenosos, como se observa en la Figura 2).

Tabla 2: promedio de emergencia de plantas de alfalfa en cada tratamiento (combinación fungicida-suelo) a partir de 300 semillas sembradas en cada maceta.

			FACTOR A (fungicida)		
			a1: con	a2: sin	media
FACTOR B (Tipo de suelo)	arcilloso	b1	228	175.33	201.66
	franco	b2	251.33	266.33	258.83
	arenoso	b3	239	271.33	255.16
	media		239.44	237.66	238.55

Efecto principal de a_1 = 239.44 - 238.55 = 0.89

Este resultado tan pequeño llevaría a correr el riesgo de concluir que **no existe un efecto debido a a_1.** Sin embargo, como se observa en la Figura 2 el factor A (fungicida) **sí tiene efectos,** que son diferentes en cada nivel del factor B (tipo de suelo), pero estos quedan enmascarados al promediarlos para calcular el efecto principal. Es por esto que **cuando existe una interacción significativa** entre factores, **carece de sentido el estudio de los efectos principales** de cada factor. En cambio interesa evaluar los **efectos simples** dentro de cada uno de los niveles del otro factor.

En el experimento de emergencia de alfalfa, los **efectos simples del nivel a_1 del factor A** (Fungicida) se pueden estimar así:

- dentro de suelo arcilloso: 228 - 201.66 = 26.34
- dentro de suelo franco-limoso: 251.33 - 258.83 = -7.5
- dentro de suelo arenoso: 239 - 255.16 = -16.16

Se observa que en los suelos arcillosos emergieron en promedio aproximadamente 26 plantas más de alfalfa en las macetas donde se aplicó el fungicida, mientras que en los suelos arenosos emergieron 16 plantas menos que el promedio en las macetas tratadas con fungicida. Es decir que el efecto del fungicida no es independiente del efecto del suelo en este caso.

En el experimento de trébol blanco, donde aparentemente los factores actúan en forma independiente (todavía no lo hemos puesto a prueba) nos interesa estudiar los efectos principales del fungicida por un lado y de los suelos por el otro y podremos extraer conclusiones acerca del efecto del fungicida aplicables indistintamente a los tres tipos de suelo y acerca del efecto de los suelos aplicables indistintamente a semillas que reciban o no la aplicación del fungicida. En cambio, en el experimento de alfalfa, donde aparentemente los factores interactúan, nos interesa estudiar el efecto de la aplicación de fungicida separadamente para cada tipo de suelo y el efecto de los suelos separadamente para las condiciones con y sin aplicación de fungicida. De aquí se deriva que la presencia de interacción entre los factores

condiciona el análisis posterior de los datos y las conclusiones del estudio. Es por esto que la primera hipótesis que se pone a prueba en los experimentos factoriales se refiere a la presencia de interacción significativa entre los factores.

La principal ventaja de los experimentos factoriales es justamente la posibilidad de evaluar si distintos factores actúan de manera independiente sobre la respuesta estudiada o si por el contrario existe interacción en las respuestas. Cuando los factores son independientes, el experimento factorial representa además un importante ahorro de tiempo y esfuerzo con respecto a la conducción de distintos experimentos unifactoriales para estudiar a cada uno de los factores por separado. Esto es así porque en ausencia de interacción, cada efecto principal se estima con la misma precisión que si todo el experimento se dedicara a ese sólo factor, dado que los efectos simples de cada factor son homogéneos a través de los niveles del otro factor. En el experimento de trébol blanco, por ejemplo, si el fungicida no interactúa con el tipo de suelo, entonces todas las repeticiones realizadas en los distintos suelos pueden ser utilizadas para estimar el efecto de la aplicación del fungicida, como si todo el experimento se hubiera destinado al estudio de ese sólo factor y al mismo tiempo para estimar el efecto de los distintos suelos, como si todo el experimento se hubiera destinado al estudio de ese sólo factor. Es decir que la eficiencia de estos experimentos se deriva de que cada observación provee al mismo tiempo información acerca de los distintos factores.

Modelo lineal para el experimento factorial

En un experimento factorial de dos factores con asignación completamente aleatoria de los tratamientos (DCA), cada observación se describe como:

$$y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

Esto significa que la **k**-ésima repetición del tratamiento que combina el nivel **i** del factor **A** y el nivel **j** del factor **B** es la suma de:

1. Una constante μ que representa la media general de todas las poblaciones;
2. El efecto principal τ_i para el Factor A a nivel i;
3. El efecto principal β_j para el Factor B a nivel j;
4. El efecto de la interacción $(\tau\beta)_{ij}$ cuando el Factor A está a nivel i y el Factor B está a nivel j.
5. El error aleatorio en la repetición k del tratamiento que combina el nivel i del factor A y el nivel j del factor B.

donde,
los ε_{ijk} son independientes y se distribuyen normalmente $\mathbf{N(0; \sigma^2)}$;
i varía desde **1** hasta **a**;
j varía desde **1** hasta **b**; y
k varía desde **1** hasta **r.**
Hay un total de **a.b.r** observaciones porque se realizan **r** repeticiones de los **a.b** tratamientos.

Del modelo se desprende que para probar si los factores interactúan debemos poner a prueba si los $(\tau\beta)_{ij}$ son significativamente distintos de cero para todos los niveles **i** y **j** del experimento o si alguno difiere de cero. Para probar si el factor A tiene un efecto significativo sobre la respuesta, podemos poner a prueba si los τ_i son significativamente distintos de cero para todos los niveles i del factor A o si alguno difiere. De manera análoga para probar el efecto del factor B postularemos una hipótesis semejante acerca de los β_j.

Si el material experimental es heterogéneo, la heterogeneidad se conoce de manera de poder controlarla y, en consecuencia, el experimento factorial de dos factores se diseña en bloques completos aleatorios (DBCA), cada observación se describe como:

$$y_{ijk} = \mu + \rho_k + \tau_i + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$

donde la única diferencia con el modelo anterior es el agregado de un término que representa el efecto del k-ésimo bloque ρ_k , y del supuesto de independencia entre bloques y tratamientos, que se debe considerar junto con los supuestos de distribución de los errores.

RESIDUALES: En un experimento factorial con diseño completamente aleatorizado, el error ε_{ijk} Se estima mediante el residual e_{ijk} como:

$$e_{ijk} = y_{ijk} - \bar{y}_{ij.}$$

Al igual que en otros tipos de experimentos, los residuales son muy útiles para estudiar la aptitud del modelo. Recordemos que antes de adoptar las conclusiones de cualquier análisis de varianza se debe probar la adecuación del modelo propuesto y que la herramienta principal para hacerlo es el análisis de los residuales.

Estimadores puntuales de los parámetros

Los estimadores de los parámetros del modelo se derivan por el método de mínimos cuadrados. Se asume para su derivación que los efectos de los factores se definen como desviaciones de la media y que por lo tanto: $\Sigma_i\tau_i = 0$, $\Sigma_j\beta_j = 0$, $\Sigma_i (\tau\beta)_{ij} = 0$ y $\Sigma_j (\tau\beta)_{ij} = 0$.

Se obtienen los estimadores que se presentan en la Tabla 3, que resultan intuitivamente interesantes, ya que la media general de todas las poblaciones se estima mediante la media general de las observaciones, los efectos principales de cada nivel de los factores τ_i y β_j se estiman mediante la diferencia entre el promedio del nivel en cuestión y la media general y el efecto de la interacción entre τ_i y β_j se estima restando al promedio de esa combinación los respectivos estimadores de los efectos de los niveles de los factores. Finalmente la k-ésima observación de la ij- ésima celda se estima mediante el promedio de las r observaciones de esa celda.

Un poco de notación

Un punto en el subíndice indica suma sobre la variable que reemplaza. Por ejemplo, la suma de las observaciones para el tratamiento correspondiente al nivel **i** del Factor **A** y el nivel **j** del Factor **B** es:

$$y_{ij.} = \sum_{k=1}^{r} y_{ijk}$$

La correspondiente media es:

$$\overline{y}_{ij.} = \frac{y_{ij.}}{r}$$

El total de las observaciones para el nivel **i** del Factor **A** es:

$$y_{i..} = \sum_{j=1}^{b} \sum_{k=1}^{r} y_{ijk}$$

Y la correspondiente media es:

$$\overline{y}_{i..} = \frac{y_{i..}}{br}$$

De modo similar, para el nivel **j** del Factor **B**, la suma de todas las observaciones y sus medias se expresan por:

$$y_{.j.} = \sum_{i=1}^{a} \sum_{k=1}^{r} y_{ijk}$$

$$\overline{y}_{.j.} = \frac{y_{.j.}}{ar}$$

Finalmente, la suma de todas las observaciones en el estudio es:

$$y_{...} = \sum_{i}^{a} \sum_{j}^{b} \sum_{k}^{r} y_{ijk}$$

y la media general es:

$$\overline{y}_{...} = \frac{y_{...}}{rab}$$

Estimación de los Efectos Principales y Simples

Aquí se presentan las fórmulas generales para el cálculo de los estimadores de los efectos para cualquier número de niveles y se ejemplifica con los datos del experimento de emergencia de alfalfa. Si bien estos estimadores no se presentan habitualmente en los resultados de un estudio, pueden ser útiles para interpretar los resultados más formales que surgen del ANOVA.

FACTOR A (fungicida): Efectos Principales

$\hat{\tau}_1 = \bar{y}_{1..} - \bar{y}_{...} = 239.44 - 238.55 = 0.89$

$\hat{\tau}_2 = \bar{y}_{2..} - \bar{y}_{...} = 237.66 - 238.55 = -0.89$

Efectos simples:

Dentro del SUELO 1 (arcilloso):

$\hat{\tau}_{1(1)} = \bar{y}_{11.} - \bar{y}_{.1.} = 228 - 201.66 = 26.34$

$\hat{\tau}_{2(1)} = \bar{y}_{21.} - \bar{y}_{.1.} = 175.33 - 201.66 = -26.34$

Dentro del SUELO 2 (francolimoso):

$\hat{\tau}_{1(2)} = \bar{y}_{12.} - \bar{y}_{.2.} = 251.33 - 258.83 = -7.5$

$\hat{\tau}_{2(2)} = \bar{y}_{22.} - \bar{y}_{.2.} = 266.33 - 258.83 = 7.5$

Dentro del SUELO 3 (arenoso):

$\hat{\tau}_{1(3)} = \bar{y}_{13.} - \bar{y}_{.3.} = 239 - 255.16 = -16.16$

$\hat{\tau}_{2(3)} = \bar{y}_{23.} - \bar{y}_{.3.} = 271.33 - 255.16 = 16.16$

FACTOR B (SUELO): Efectos principales

$\hat{\beta}_1 = \bar{y}_{.1.} - \bar{y}_{...} = 201.66 - 238.55 = -36.89$

$\hat{\beta}_2 = \bar{y}_{.2.} - \bar{y}_{...} = 258.83 - 238.55 = 20.28$

$\hat{\beta}_3 = \bar{y}_{.3.} - \bar{y}_{...} = 255.16 - 238.55 = 16.61$

Efectos simples

Dentro de fungicida 1 (con aplicación)

$\hat{\beta}_{1(1)} = \bar{y}_{11.} - \bar{y}_{1..} = 228 - 239.44 = -11.44$

$\hat{\beta}_{2(1)} = \bar{y}_{12.} - \bar{y}_{1..} = 251.33 - 239.44 = 11.89$

$\hat{\beta}_{3(1)} = \bar{y}_{13.} - \bar{y}_{1..} = 239 - 239.44 = -0.44$

Dentro de fungicida 2 (sin aplicación)

$\hat{\beta}_{1(2)} = \bar{y}_{21.} - \bar{y}_{2..} = 175.33 - 237.66 = -62.33$

$\hat{\beta}_{2(2)} = \bar{y}_{22.} - \bar{y}_{2..} = 266.33 - 237.66 = 28.67$

$\hat{\beta}_{3(2)} = \bar{y}_{23.} - \bar{y}_{2..} = 271.33 - 237.66 = 33.67$

Tabla 3: Estimadores de mínimos cuadrados para los parámetros del modelo factorial

Parámetro	*Estimador*
$\mu_{..}$	$\bar{y}_{...}$
τ_i	$\bar{y}_{i..} - \bar{y}_{...}$
β_j	$\bar{y}_{.j.} - \bar{y}_{...}$
$(\tau\beta)_{ij}$	$\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}$
$\mu_{ij} = \mu_{..} + \tau_i + \beta_j + (\tau\beta)_{ij}$	$\bar{y}_{ij.}$

Análisis de la varianza para un experimento factorial

El análisis de la varianza de los experimentos factoriales comprende inicialmente la misma descomposición que ya presentamos para experimentos de un solo factor de la suma de cuadrados de las desviaciones de cada observación y_{ijk} con respecto a la media general $\bar{y}_{...}$ en las componentes debidas a los tratamientos y al error experimental (si se trata de un diseño en bloques, lógicamente también se incluye la componente debida a los bloques). El aspecto particular de los experimentos factoriales reside en la posibilidad de ***particionar la suma de cuadrados de tratamientos*** en componentes atribuibles a los distintos factores y a la interacción entre ellos.

Primero, vamos a presentar la descomposición de la desviación total $y_{ijk} - \bar{y}_{...}$ considerando al estudio como con **a.b** tratamientos, que no difiere de la descomposición presentada en capítulos anteriores:

$$\underbrace{y_{ijk} - \bar{y}_{...}}_{\text{Desviación Total}} = \underbrace{\bar{y}_{ij.} - \bar{y}_{...}}_{\substack{\text{Desviación de la media del tratamiento alrededor de la media general}}} + \underbrace{y_{ijk} - \bar{y}_{ij.}}_{\substack{\text{Desviación alrededor de la media del tratamiento}}}$$

$$SC_{TOTAL} = SC_{TRAT} + SC_E$$

donde:

$$SC_{TOTAL} = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{...})^2$$

$$SC_{TRAT} = r \sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{...})^2$$

$$SC_E = \sum_i \sum_j \sum_k (y_{ijk} - \bar{y}_{ij.})^2$$

La SCTRAT refleja la variación entre promedios de los **ab** tratamientos y es la suma de cuadrados de tratamientos, y la SCE refleja la variación dentro de tratamientos y es la suma de cuadrados del error experimental. Nótese que hasta aquí la única diferencia con la partición presentada para experimentos de un solo factor es la notación utilizada para designar a cada tratamiento, ya que ahora se usan dos subíndices i y j para identificar a los niveles de los dos factores.

Partición de la SCTRAT

La particularidad que ofrece la estructura factorial de los tratamientos es la posibilidad de descomponer la suma de cuadrados de tratamientos en términos de componentes que reflejan el efecto principal de A, el efecto principal de B, y la interacción AB.

$$\underbrace{\bar{y}_{ij.} - \bar{y}_{...}}_{\text{Desviación de la media de tratamiento a la media general}} = \underbrace{\bar{y}_{i..} - \bar{y}_{...}}_{\text{Efecto principal de A}} + \underbrace{\bar{y}_{.j.} - \bar{y}_{...}}_{\text{Efecto principal de B}} + \underbrace{\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...}}_{\text{Efecto de Interacción AB}}$$

Las sumas de los cuadrados de estos desvíos conforman la siguiente partición

$$SC_{TRAT} = SC_A + SC_B + SC_{AB}$$

donde

$$SC_A = rb\sum_i (\bar{y}_{i...} - \bar{y}_{...})^2$$

$$SC_B = ra\sum_j (\bar{y}_{.j.} - \bar{y}_{...})^2$$

$$SC_{A*B} = r\sum_i \sum_j (\bar{y}_{ij.} - \bar{y}_{i..} - \bar{y}_{.j.} + \bar{y}_{...})^2$$

SCA (suma de cuadrados del factor A) mide la variación de las medias estimadas de los niveles del factor A. Cuanto más variables sean, mayor será la SCA. Similarmente SCB (suma de cuadrados del factor B) mide la variación de las medias estimadas de los niveles del factor B. Finalmente, la SCAB es la suma de cuadrados de las interacciones estimadas para los tratamientos ab.

Partición de los grados de libertad

Correspondiente a la partición de la suma de cuadrados de tratamientos podemos, también, obtener una partición de los **ab-1** grados de libertad asociados, SCA tiene **a-1** grados de libertad asociados con ella. Hay **a** medias de niveles del factor y están sujetas a la restricción $\Sigma(\bar{y}_{i..} - \bar{y}_{...}) = 0$. Similarmente, SCB tiene **b-1** grados de libertad asociados con ella. Los grados de libertad asociados con la SCAB (la suma de cuadrados de la interacción) se calculan como: (ab-1)-(a-1)-(b-1) = **(a-1) (b-1)**

Cuadrados medios

Los cuadrados medios se obtienen dividiendo las sumas de cuadrados por sus grados de libertad asociados. La siguiente tabla resume el ANOVA de un experimento factorial.

Fuente de Variación	GL	SC	CM	Valor esperado de CM
Factor A	a-1	SC_A	SC_A/GL_A	$\sigma^2_E+rb\sum\tau_i^2/(a-1)$
Factor B	b-1	SC_B	SC_B/GL_B	$\sigma^2_E+ra\sum\beta_i^2/(b-1)$
Interacción (AxB)	(a-1)(b-1)	SC_{AB}	SC_{AB}/GL_{AB}	$\sigma^2_E+r\sum(\tau\beta)_{ij}^2/(a-1)(b-1)$
Error Experimental	ab(r –1)	SC_{EE}	SC_{EE}/GL_{EE}	σ^2_E
Total	abr-1 = n-1	SC_T		

En la última columna de la tabla se presentan las funciones de los parámetros poblacionales que son estimadas por los cuadrados medios. Se observa que si son ciertas las hipótesis nulas que postulan que los efectos τ_i , β_j , y $(\tau\beta)_{ij}$ son todos ceros, entonces todos los CM estiman a la misma varianza σ^2_E. . En cambio, si los efectos son importantes, los respectivos CM serán mayores que el CMerror. En consecuencia, para probar el significado de los parámetros se calcula el cociente F que divide al respectivo CM por el CMerror y valores grandes de estos cocientes indican que los datos no concuerdan con las hipótesis nulas.

Pruebas de hipótesis

Las pruebas de hipótesis acerca de los parámetros del modelo se basan en los cocientes apropiados de los cuadrados medios. Estos cocientes siguen una distribución F con los grados de libertad correspondientes al numerador y denominador del cociente.

Normalmente, el análisis del experimento factorial comienza con una prueba que determina si los factores interactúan. Sólo si no se encuentran evidencias suficientes para suponer la presencia de interacción se conducen las pruebas para los efectos principales de los factores.

Prueba para el efecto de interacción

H_0: $(\tau\beta)_{ij} = 0$, para todo i y j
H_1: Algún $(\tau\beta)_{ij}$ distinto de cero

La prueba estadística apropiada es:

$$F^* = CM(A*B)/CMEE$$

Grandes valores de F^* indican la existencia de interacciones significativas. Cuando H_0 es cierta, F^* se distribuye como $F_{((a-1)(b-1);glee)}$, donde glee son los grados de libertad del error experimental. Así, la regla de decisión apropiada para controlar el Error de tipo I en α es:

Si $F^* \leq F_{(1-\alpha;\,(a-1)(b-1);glee)}$, no hay evidencias para rechazar H_0
Si $F^* > F_{(1-\alpha;\,(a-1)(b-1);glee)}$, se rechaza H_0 y se concluye según H_1

Donde $F_{(1-\alpha;(a-1)(b-1);\,glee)}$ es el $(1-\alpha)100$ percentil de la distribución F apropiada.

Prueba para los efectos del factor A

Las pruebas para los efectos principales de los factores A y B, normalmente, siguen a la prueba para las interacciones, pero sólo se llevan a cabo cuando no hay evidencias de interacción significativa entre los factores. Para probar si es significativo el efecto principal del factor A:

H_0: $\tau_1 = \tau_2 = ... = \tau_a = 0$
H_1: algún τ_i distinto de cero

Utilizamos la prueba estadística:

$F^* = CMA/CMEE$

Nuevamente, grandes valores de F^* indican la existencia de efecto principal de A. Ya que F^* se distribuye como $F_{(a-1;glee)}$ cuando H_0 es cierta, la regla de decisión apropiada para controlar el riesgo de cometer error de tipo I con nivel de significación α es:

Si $F^* \leq F_{(1-\alpha;\,(a-1);glee)}$, no hay evidencias para rechazar H_0
Si $F^* > F_{(1-\alpha;\,(a-1);glee)}$, se rechaza H_0 y se concluye según H_1

Prueba para los efectos del factor B

Esta prueba es similar a la de los efectos del factor A. Las alternativas son:

H_0: $\beta_1 = \beta_2 = ... = \beta_b = 0$
H_1: algún β_j es distinto de cero

La prueba estadística es:

$F^* = CMB/CMEE$

y la regla de decisión apropiada para controlar el riesgo de cometer error de tipo I en α es:

Si $F^* \leq F_{(1-\alpha;\,(b-1);glee)}$, no hay evidencias para rechazar H_0
Si $F^* > F_{(1-\alpha;\,(b-1);glee)}$, se rechaza H_0 y se concluye según H_1

Los programas estadísticos informan el valor P: la probabilidad de $F_{(GL\ factor,\ GL\ error)} > F^*$, con esta información la regla de decisión equivalente a las presentadas es: **si el valor P < α, se rechaza H_0 con nivel de significación α.**

Ahora estamos en condiciones de poner a prueba las hipótesis que nos permitirán contestar las preguntas asociadas con los experimentos de emergencia de las leguminosas forrajeras que fueron expuestos al comienzo del capítulo y en las Figuras 1 y 2. Presentamos a continuación para los dos experimentos las tablas de ANOVA (Tabla 4 para trébol blanco y Tabla 5 para alfalfa), las pruebas de interés y las conclusiones.

Tabla 4: Análisis de la Varianza para emergencia de Trébol Blanco
(los promedios se presentaron en la Figura 1)

F.V.	SC	gl	CM	F	valor P
Fungicida	1568,00	1	1568,00	3,42	0,0890
Suelo	18201,00	2	9100,50	19,88	0,0002
Fungicida*Suelo	39,00	2	19,50	0,04	0,9584
Error	5494,00	12	457,83		
Total	25302,00	17			

Primero ponemos a prueba el efecto de **interacción fungicida x suelo**
H_0: $(\tau\beta)_{ij} = 0$, para todo i y j i=1, 2 j= 1, 2, 3
H_1: Algún $(\tau\beta)_{ij}$ distinto de cero

La prueba estadística apropiada es:

$F^* = CMA*B/CMEE = 19.50/457.83 = 0.04$

Se observa que el valor-p asociado al estadístico F de la interacción (Fungicida x Suelo) es > α=0,05 lo que indica que estos datos no aportan evidencia suficiente para afirmar que los efectos del fungicida y el tipo de suelo sobre la emergencia de trébol blanco interactúan significativamente. No hay evidencias para rechazar H_0

Entonces se pueden estudiar separadamente los efectos de cada uno de los factores.

Para el **efecto del fungicida**:
H_0: $\tau_1 = \tau_2 = 0$
H_1: algún τ_i distinto de cero i=1, 2

$F^* = CMFung/CMEE = 1568/457.83 = 3.42$

El valor P asociado es 0.089, un valor marginal que no permite descartar la posibilidad de que el fungicida tenga efecto sobre la emergencia de trebol blanco.

Para el **efecto del tipo de suelo**:

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$
H_1: algún β_j es distinto de cero $\quad j=1, 2, 3$

La prueba estadística es:

$$F^* = CMB/CMEE = 9100.50/ 457.83 = 19.88$$

El valor P asociado a esta prueba es $0.0002 < \alpha = 0.05$, es decir que el experimento provee evidencias para afirmar que la emergencia promedio de trébol difiere entre tipos de suelo más de lo esperable por mero azar. Se rechaza H_0

Se concluye que al menos alguno de los suelos presenta efecto significativamente distinto y se procede a realizar el test de Tukey para investigar cuáles son los suelos que presentan diferencias significativas.

Tabla 4: Test de Tukey para la comparación de medias de emergencia de trébol blanco en distintos suelos. Previamente en el ANOVA no se detectó interacción significativa y se rechazó la hipótesis de igual emergencia de plantas en los distintos suelos.

Test:Tukey Alfa:=0.05 DMS:=32.95831
Error: 457.8333 gl: 12

Suelo	Medias	n		
arcilloso	119.33	6	A	
francolimo	164.83	6		B
arenoso	196.83	6		B

Letras distintas indican diferencias significativas(p<= 0.05)

Se concluye que la emergencia es significativamente mayor en suelos de texturas medias a gruesas: francolimosos y arenosos, que no difieren entre sí y en cambio, presentan emergencia promedio significativamente mayor que los suelos arcillosos.

Para el experimento de emergencia de alfalfa, se presentan los resultados del ANOVA en la Tabla 5. En este caso, el valor-p para la interacción fungicida x suelo es <0,0001 lo que indica que existen efectos de interacción, es decir que el fungicida no produce el mismo efecto sobre la emergencia en los diferentes suelos y que la respuesta de la emergencia a los tipos de suelo difiere si se aplica o no el fungicida. Como consecuencia, carece de sentido estudiar los efectos principales de fungicida y suelo.

Tabla 5: Análisis de la Varianza para emergencia de Alfalfa (Los promedios se presentaron en la Figura 2)

F.V.	SC	gl	CM	F	p-valor
Fungicida	12,50	1	12,50	1,84	0,1994
Suelo	12254,33	2	6127,17	904,01	<0,0001
Fungicida*Suelo	6086,33	2	3043,17	448,99	<0,0001
Error	81,33	12	6,78		
Total	18434,50	17			

Entonces deben compararse las medias de los distintos tratamientos usando la prueba de Tukey para la interacción (Tabla 6).

Tabla 6: Prueba de Tukey para la comparación de medias de tratamientos (combinaciones de los niveles de los factores), previamente se detectó interacción significativa en el ANOVA (Tabla 5)

Test:Tukey Alfa:=0.05
DMS:=7.141
Error: 6.7778 gl: 12

Fungicida	Suelo	Medias	n					
sin	arcilloso	175,33	3	A				
con	arcilloso	228	3		B			
con	arenoso	238,67	3			C		
con	francolimo	251,33	3				D	
sin	francolimo	266,33	3					E
sin	arenoso	271,33	3					E

Letras distintas indican diferencias significativas
(p<= 0.05)

De acuerdo a estos resultados, para obtener al momento de la emergencia un número promedio de plantas de alfalfa significativamente mayor, se recomienda sembrar en suelos de texturas medias a gruesas (indistintamente en suelos francolimosos o en suelos arenosos), sin aplicación de fungicida, ya que los tratamientos sin fung.+suelo arenoso y sin fun..+suelo francolimoso no difieren estadísticamente entre sí en cuanto al número medio de plantas que emergen y en cambio, difieren de todos los otros tratamientos ensayados. También se concluye que si se siembra en suelos arcillosos conviene aplicar fungicida ya que sin fun.+suelo arcilloso evidenció diferencias significativas respecto a con fun.+suelo arcilloso.

EXPERIMENTOS FACTORIALES Y CIENCIAS AGROPECUARIAS

Los experimentos factoriales se utilizan prácticamente en todos los campos de aplicación. Son de gran valor en trabajos exploratorios cuando se sabe poco sobre el nivel óptimo de los factores, o si ni siquiera se conoce cuáles son factores importantes. Por ejemplo cuando se está investigando un cultivo nuevo y se deben establecer prácticas culturales, ya sea porque la especie a introducir nunca fue cultivada comercialmente o porque sólo fue cultivada en ambientes diferentes y se deben seleccionar las pautas óptimas de manejo para las condiciones locales. A continuación se presenta un experimento factorial de este tipo y se muestra la secuencia que sigue el análisis de los datos.

EJEMPLO 1: PRODUCCIÓN DE HONGOS COMESTIBLES

Se realizó un experimento para estudiar el efecto de la cepa y el sustrato en la producción de un hongo comestible conocido como Gírgola (*Pleorotus ostratus*). La unidad experimental fue cada bolsa sembrada con el hongo. Las bolsas con las combinaciones de cepa y sustrato fueron dispuestas al azar en la cámara de cultivo. Se emplearon 4 bolsas por tratamiento, evaluándose al final del período de cultivo el rendimiento en kg/bolsa. La Tabla 7 contiene los resultados del estudio. Allí se puede observar que, por ejemplo, la observación correspondiente a la tercera bolsa de la Cepa 2 y Sustrato 3 (y_{233}) es 2.02 kg., y la media del tratamiento correspondiente a la combinación Cepa 2 y Sustrato 3 ($\bar{y}_{23.}$) es 2.005.

Tabla 7: Evaluación del rendimiento en kg/bolsa de 3 cepas de *Pleorotus ostreatus* (1, 2, 3) con tres sustratos

CEPAS (FACTOR A)	SUSTRATOS (FACTOR B)			Promedio de cepas	
	1 Paja de trigo + aserrín de álamo	2 Paja de alfalfa + aserrín de álamo	3 Paja de trigo	Total ($y_{i..}$)	Media ($\bar{y}_{i..}$)
1 (1 Repeticiones)	1.92 2.12 2.012 1.82 $\bar{y}_{11.}$ 2.19	1.52 1.61 1.525 1.48 $\bar{y}_{12.}$ 1.49	2.03 2.10 2.145 2.25 $\bar{y}_{13.}$ 2.20	22.73 ($y_{1..}$)	1.894
2	1.88 1.95 1.883 1.99 $\bar{y}_{21.}$ 1.71	1.39 0.98 1.325 1.41 $\bar{y}_{22.}$ 1.52	1.93 1.97 2.005 2.02 $\bar{y}_{23.}$ 2.10	20.85 ($y_{2..}$)	1.737
3	2.02 1.94 2.025 2.25 $\bar{y}_{31.}$ 1.89	0.82 1.72 1.503 1.68 $\bar{y}_{32.}$ 1.79	2.26 1.98 2.110 2.23 $\bar{y}_{33.}$ 1.97	22.55 ($y_{3..}$)	1.879
Total ($y_{.j.}$)	23.68 ($y_{.1.}$)	17.41 ($y_{.2.}$)	25.04 ($y_{.3.}$)	66.13 ($y_{...}$)	
Media ($\bar{y}_{.j.}$)	1.973	1.451	2.087	1.837	

Vamos a investigar la presencia de efectos de cepas y sustratos en la producción de gírgolas, usando un nivel de significación de 0.05 para cada prueba. Primero, comenzamos probando si hay o no efectos de interacción:

H_0: todos los $(\tau\beta)_{ij} = 0$ i=1, 2, 3 j=1, 2, 3

H_1: No todos los $(\tau\beta)_{ij}$ son 0

Utilizando los datos de la Tabla 7 para la prueba estadística:

$F^*=0.002236 / 0.042064 = 0.053157$

Ya que vamos a controlar el riesgo de cometer error de Tipo I en 0.05, requerimos F(0.95, 4, 27), en nuestras tablas debemos entrar con 0.05, la probabilidad de encontrar valores mayores al punto crítico. No encontramos el valor exacto de los grados de libertad del error experimental por lo que podemos buscar el valor de F con grados de libertad aproximados, en este caso buscamos con 24 grados de libertad, siendo el valor de F = 2.776. De modo que la regla de decisión es:

Si $F^* \leq 2.776$ concluya H_0

Si $F^* > 2.776$ concluya H_1

Ya que F^* es 0.053157, concluimos H_0, que no hay evidencias para afirmar que la cepa y el sustrato interactúan en sus efectos sobre la producción.

Luego probamos el efecto principal de la cepa (Factor A).

H_0: $\tau_1 = \tau_2 = \tau_3 = 0$

H_1: algún τ_i distinto de cero

$F^* = 0.089678 / 0.042064 = 2.131942$

Para el nivel de significación de 0.05 requerimos F(0.95, 2, 27), aproximando los grados de libertad del error experimental a 24, F(0.95, 2, 24) = 3.403. Ya que $F^* = 2.131942 < 3.403$, concluimos según H_0, que todas las medias $\mu_{i.}$ son iguales, o que la cepa no tiene efecto en la producción del hongo.

Finalmente, probamos los efectos principales del sustrato (Factor B):

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

H_1: algún β_j es distinto de cero

$F^* = 1.380269 / 0.042064 = 32.813546$

Para $\alpha = 0.05$ y aproximando los grados de libertad del error experimental a 24 F(0.95,2,24)= 3.403. Ya que $F^* = 32.813546 > 3.403$, concluimos según H_1, que no todas las medias de los niveles del factor B $\mu_{.j}$ son iguales, o que existen ciertos efectos asociados con el sustrato.

En la Tabla 8 se presentan los resultados del ANOVA. No es necesario utilizar las tablas de F ya que el software provee los valores p para las distintas pruebas. Así vemos, que el valor p para la prueba de la interacción es 0.9944 > 0.05, lo que nos lleva a la misma conclusión que llegamos anteriormente, buscando el valor crítico de la distribución de F. Sólo la prueba correspondiente a la hipótesis de nulidad de efectos del factor B presenta un valor p menor a 0.05, indicando que el valor F^* es demasiado grande como para aceptar H_0, por lo que la rechazamos.

Tabla 8: Análisis de Varianza para el experimento de producción de hongos comestibles.

FV	SC	GL	CM	F	Valor-p
CEPAS (A)	0.179356	2	0.089678	2.13	0.138144
SUSTRATO (B)	2.760538	2	1.380269	32.81	0.000000
INTERACCION (A*B)	0.008944	4	0.002236	0.05	0.994400*
ERROR EXP.	1.135728	27	0.042064		
TOTAL	4.084566	35			

Estas pruebas muestran que sólo el sustrato tiene un efecto significativo sobre el rendimiento de las gírgolas entre los tratamientos estudiados. En este punto sería claramente deseable llegar a seleccionar el o los sustratos que producen los mayores rendimientos. Podemos comparar las tres medias de sustratos mediante una prueba de Tukey. Como llegamos a la conclusión que los factores no interactúan, y por lo tanto sabemos que los efectos de los sustratos son constantes para todas las cepas, comparamos las medias marginales de sustratos. La diferencia mínima significativa según Tukey será:

$$\Delta TUK = t_{TUK\,(b,glee,0.05)}\sqrt{\frac{2(CMEE)}{ar}}$$

En este experimento, y aproximando los grados de libertad del error experimental a 24, tenemos que:

$$\Delta TUK = 2.50\sqrt{\frac{2(0.042064)}{(3)(4)}} = 0.209$$

De modo que vamos a considerar como significativa toda diferencia entre medias muestrales mayor a 0.209 kg. En la Tabla 9 se presentan los promedios ordenados y las diferencias absolutas que se comparan con el valor DELTA.

Tabla 9: Comparación de medias de rendimiento de Gírgolas en distintos sustratos, posterior al ANOVA que no detectó efecto significativo de interacción y en cambio encontró diferencias significativas en el efecto principal de los sustratos.

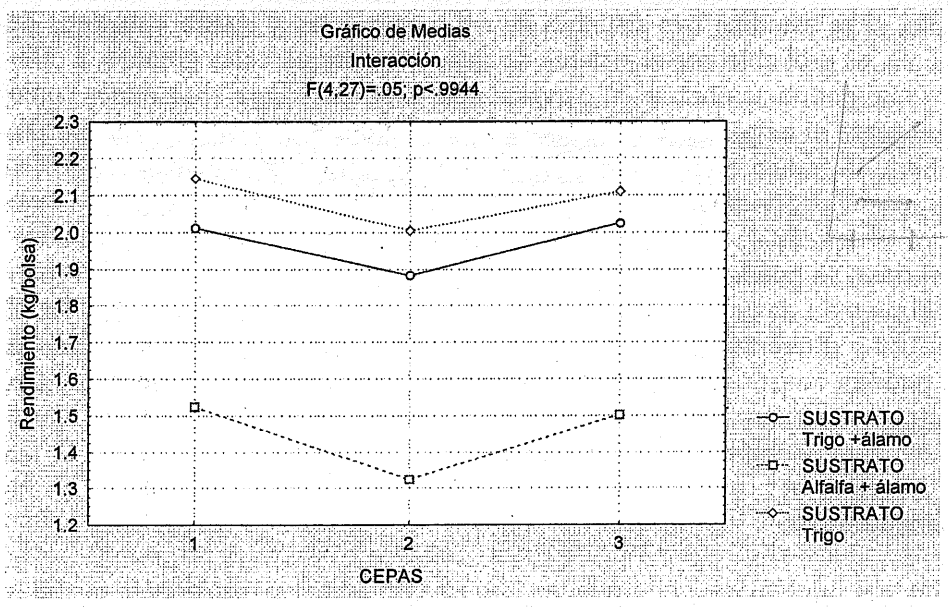
SUSTRATOS	Medias	Paja de trigo 2.087	Paja de trigo + aserrín de álamo 1.973	Paja de alfalfa + aserrín de álamo 1.451
P. de alfalfa + aserrín de álamo	1.451	0.636 *	0.522 *	-------
P. de trigo + aserrín de álamo	1.973	0.114	--------	
P. de trigo	2.087	--------		

Podemos observar que dos diferencias entre sustratos superan el valor mínimo significativo. La media del sustrato con alfalfa es significativamente menor a las de los otros sustratos, siendo no significativa la diferencia de las medias de los sustratos que contienen paja de trigo.

Se puede concluir que para las condiciones en que se llevó a cabo el experimento, cualquier cepa produce el mismo rendimiento y los sustratos con paja de trigo ensayados superan en rendimiento promedio al de paja de alfalfa y aserrín de álamo.

Sin lugar a dudas, los gráficos de medias son de gran utilidad especialmente en los estudios con más de un factor. En los gráficos de medias de tratamientos podemos visualizar la existencia de posibles interacciones, además nos facilitan la interpretación de los resultados. La Figura 3 presenta las medias de las nueve combinaciones de los niveles de factores para este experimento.

Figura 3: Medias de rendimiento de distintas cepas de gírgolas creciendo en diferentes sustratos.



Puede existir interacción

Se puede observar como las diferencias entre sustratos se mantienen casi constantes en todas las cepas, haciendo que las líneas que unen los puntos sean prácticamente paralelas. El sustrato con alfalfa fue inferior en rendimiento medio a los otros sustratos por una diferencia similar en todas las cepas. La diferencia que se observa entre las líneas que unen los puntos de las medias de los otros dos sustratos puede considerarse como debida a las fluctuaciones del muestreo teniendo en cuenta los resultados de la prueba de Tukey. Del mismo modo, las diferencias entre las medias de las cepas para los distintos sustratos observadas en el gráfico no llegan a ser significativas en el análisis de varianza, por lo que podemos considerarlas como debidas a las fluctuaciones en el muestreo.

DISEÑO EN PARCELAS DIVIDIDAS

En los experimentos factoriales hasta aquí presentados, los tratamientos estan constituidos por combinaciones de niveles de factores de igual jerarquía y se asignan al azar a las unidades experimentales de acuerdo al diseño seleccionado: DCA o DBCA por ejemplo. En cambio existen situaciones en las que uno de los factores requiere unidades experimentales de mayor tamaño que el otro y también proyectos que requieren detectar diferencias con mayor precisión en un factor que en los demás. Para esos casos se recomienda diseñar experimentos de parcelas divididas. En este diseño existen dos tamaños de unidades experimentales. En las parcelas grandes se asignan por ejemplo los niveles del factor que requiere unidades experimentales de mayor tamaño. Estas parcelas se dividen, en tantas subparcelas como niveles del otro factor se presenten en el experimento. De este modo, el proceso de aleatorización se cumple en dos etapas. La primera etapa corresponde a la asignación de los

niveles del primer factor según el diseño de base, DCA o DBCA, y en la segunda etapa se asignan al azar los niveles del otro factor en las subparcelas dentro de cada parcela grande.

Por ejemplo, si se estudian tres sistemas de riego diferentes y cuatro cultivares de una especie hortícola, es probable que el sistema de riego requiera parcelas grandes, a diferencia del cultivar que podría sembrarse en parcelas más pequeñas. Para situaciones como la planteada es frecuente el empleo en agricultura del diseño en parcelas divididas (DPD), que, como su nombre lo indica, tiene su origen en los ensayos con parcelas. Los posibles usos del diseño en parcelas divididas no se limitan a la agricultura. En un ensayo para evaluar el efecto sobre la terneza de tres tipos de ración alimenticia y dos métodos de conservación de carne vacuna, se puede pensar en este diseño asignando la ración al animal (parcela completa) y el método de conservación a cada media res (subparcela). Si se estudia el efecto de temperatura y diferentes medios de cultivo en el desarrollo de un hongo, las cámaras de cultivo podrían constituir las parcelas completas y los medios de cultivo las subparcelas. Asimismo, en la industria, se pueden replicar mezclas de materias primas, que a su vez se subdividen para recibir distintos tipos de curado.

Dos errores experimentales

Para la formulación del modelo para el diseño en parcelas divididas se consideran distintas varianzas del error experimental para la subparcela y la parcela completa, lo que incluye los efectos del error aleatorio por separado para ambas. La consideración de dos errores separados es consecuencia del hecho de que las observaciones de distintas subparcelas en la misma parcela pueden tener correlación positiva. La correlación refleja la naturaleza de las unidades experimentales que responden de modo similar cuando son adyacentes, como las subparcelas vecinas, las medias reces de un mismo animal o los medios de cultivo dentro de una misma cámara. Se supone que existe una correlación ρ entre las observaciones en cualesquiera dos subparcelas en la misma parcela completa y que las observaciones de dos parcelas completas diferentes no están correlacionadas. Dados estos supuestos, se puede demostrar que la varianza del error para los efectos principales del factor asignado a las parcelas completas, que llamaremos factor A es $\sigma^2[1+(b-1)\rho]$ si hay b subparcelas por cada parcela completa. Del mismo modo, para las comparaciones entre los efectos principales del factor B y las de la interacción AB la varianza del error por cada subparcela es $\sigma^2(1-\rho)$. Vemos que si el coeficiente de correlación es positivo la varianza del error para los efectos del factor A va a ser mayor que la correspondiente para las comparaciones de los efectos principales del factor B y la interacción.

Teniendo en cuenta estas características de las varianzas de los errores y el empleo de unidades experimentales grandes que contienen a las subunidades se aconseja el diseño en parcelas divididas en las siguientes situaciones:

- Cuando los niveles de uno o más factores necesitan mayores cantidades de material experimental en una unidad experimental que los niveles de otros factores.

- Cuando se incorpora en un experimento un factor adicional para aumentar su alcance. Por ejemplo, si el objetivo del experimento es comparar fungicidas como protectores contra infección por una enfermedad, se puede aumentar el alcance de las con-

clusiones incluyendo varias variedades de las cuales se conoce que difieren en su resistencia a la enfermedad.

- Cuando se sabe que pueden esperarse diferencias mayores entre los niveles de ciertos factores que entre niveles de otros. En este caso, los niveles del factor o combinación de niveles de los factores en los que se esperan diferencias grandes pueden asignarse aleatoriamente a las unidades grandes, simplemente por comodidad.

- Cuando se desea mayor precisión para comparaciones entre ciertos factores, que para otras.

Como consecuencia de las diferencias en los errores asociados con las comparaciones de parcelas completas y subparcelas, la partición de la suma de cuadrados en el análisis de varianza presenta algunas diferencias con respecto a la partición en un experimento multifactorial sin parcelas divididas como el que vimos en un capítulo anterior. Las particiones para los efectos del factor y los bloques permanecen igual que los diseños multifactoriales usuales, pero el error experimental se divide en dos componentes. Una componente del error se asocia con el factor de la parcela completa y la otra con el factor de la subparcela y la interacción.

El Modelo

Si el factor correspondiente a la parcela completa se asigna según un diseño en bloques completos aleatorizados el modelo lineal es:

$$y_{ijk} = \mu + \tau_i + \rho_k + \delta_{ik} + \beta_j + (\tau\beta)_{ij} + \varepsilon_{ijk}$$
$$i = 1,2,...,a \qquad j = 1,2,...,b \qquad k = 1,2,...,r$$

donde μ es la media general, τ_i es el efecto del i-ésimo nivel del factor A, ρ_k es el efecto del k-ésimo bloque, δ_{ik} es el error aleatorio de la parcela completa, β_j es el efecto del j-ésimo nivel del factor B, $(\tau\beta)_{ij}$ es el efecto de interacción entre ambos factores y ε_{ijk} es el error aleatorio de la subparcela.

Se supone que los errores de la parcela completa y la subparcela son errores aleatorios independientes con distribución normal con media 0 y varianzas respectivas σ_δ^2 y σ_ε^2. La asignación aleatoria de los niveles de los factores a las unidades experimentales justifica el supuesto de independencia para los dos errores aleatorios y la correlación igual entre los errores de las unidades en las subparcelas dentro de una misma parcela completa.

Análisis de Varianza

A continuación se presenta el análisis de varianza y las esperanzas de los cuadrados medios correspondientes al modelo del diseño en parcelas divididas en bloques:

FV	GL	CM	E(CM)
Bloques	r-1	CMBloques	
A	a-1	CMA	$\sigma_e^2 + b\sigma_d^2 + rb\theta_a^2$
Error(a)	(a-1)(r-1)	CME(a)	$\sigma_e^2 + b\sigma_d^2$
B	b-1	CMB	$\sigma_e^2 + ra\theta_b^2$
A*B	(a-1)(b-1)	CMA*B	$\sigma_e^2 + r\theta_{ab}^2$
Error(b)	a(r-1)(b-1)	CME(b)	σ_e^2

Las esperanzas de los cuadrados medios para el Error(a) y Error(b) reflejan la diferencia en la variabilidad de los dos tipos diferentes de unidades experimentales.

Errores estándar en el diseño en parcelas divididas

Los errores estándar para las pruebas de hipótesis para las comparaciones entre medias se presentan en la siguiente tabla:

Comparación de medias	Error estándar estimado
Diferencia entre niveles del factor A $\bar{y}_{i..} - \bar{y}_{i'..}$	$\sqrt{\frac{2CME(a)}{rb}}$
Diferencia entre niveles del factor B $\bar{y}_{.j.} - \bar{y}_{.j'.}$	$\sqrt{\frac{2CME(b)}{ra}}$
Diferencia entre dos niveles de B a un mismo nivel de A $\bar{y}_{ij.} - \bar{y}_{ij'.}$	$\sqrt{\frac{2CME(b)}{r}}$
Diferencia entre dos niveles de A al mismo o distinto nivel de B $\bar{y}_{ij.} - \bar{y}_{i'j.}$ o $\bar{y}_{ij.} - \bar{y}_{i'j'.}$	$\sqrt{\frac{2[(b-1)CME(b) + CME(a)]}{rb}}$

Los grados de libertad asociados con cada uno de los errores estándar son los correspondientes a los cuadrados medios empleados en el error estándar. La única excepción es la comparación final en la que el error estándar es una combinación ponderada de los dos cuadrados medios de error. En este caso, los grados de libertad apropiados pueden ser aproximados por el procedimiento introducido por Satterwaite del siguiente modo:

$$gl = \frac{[(b-1)CME(b) + CME(a)]^2}{\frac{[(b-1)CME(b)]^2}{gle(b)} + \frac{[CME(a)]^2}{gle(a)}}$$

donde gle(a) y gle(b) son los grados de libertad correspondientes a los CME(a) y CME(b), respectivamente, del análisis de varianza.

Ejemplo de experimento con Diseño en Parcelas Divididas

Se condujo un experimento para evaluar el efecto de tres métodos de inoculación de bacterias aplicadas a dos cultivares de trébol en el rendimiento en peso seco en kg/parcela (PS). El experimento fue conducido en 4 bloques, cada bloque fue dividido por la mitad, asignándose aleatoriamente un cultivar a cada mitad. Cada mitad fue subdividida en 3 parcelas, a las que se les asignó aleatoriamente los tres métodos de inoculación. Los datos se presentan a continuación:

BLOQUE	CULTIVAR	INOC	PS
1	1	1	27.4
1	1	2	29.7
1	1	3	34.5
1	2	1	34.4
1	2	2	37.5
1	2	3	39.4
2	1	1	28.9
2	1	2	28.7
2	1	3	33.4
2	2	1	33.7
2	2	2	37.4
2	b	3	41.4
3	1	1	28.6
3	1	2	29.7
3	1	3	32.9
3	2	1	32
3	2	2	34.7
3	2	3	37.6
4	1	1	26.7
4	1	2	28.9
4	1	3	31.8
4	2	1	31
4	2	2	33.5
4	2	3	35.7

H_0: Todos los $(\tau\beta)_{ij}= 0$
H_1: Algún $(\tau\beta)_{ij}\neq 0$

F=CMA*B/CME(b)

En este ejemplo F= 1.43/ 0.58 = 2.47 corresponde a un valor p de 0.1265 en la distribución F(2;12). La hipótesis de no interacción no puede ser rechazada, por lo que proseguimos a analizar los efectos principales de los factores.

H_0: Todos los $\tau_i= 0$
H_1: Algún $\tau_i\neq 0$

F=CMA /CME(a)

En el ejemplo F= 187.6/ 3.34 = 56.1 corresponde a un valor p de 0.0049 en la distribución F(1;3). En este caso se rechaza la hipótesis nula. La diferencia entre las medias de los cultivares es significativa. Siendo sólo dos los niveles de este factor no se requieren pruebas posteriores. El cultivar 2 presentó un peso seco promedio significativamente superior.

H_0: Todos los $\beta_j = 0$
H_1: Algún $\beta_j \neq 0$

F=CMB/CME(b)

Vemos que F= 61.38/ 0.58 = 105.83 corresponde a un valor p < 0.0001 en la distribución F(2;12). En este caso se rechaza la hipótesis nula. La diferencia entre las medias de los cultivares es significativa. La prueba posterior de Tukey nos permite concluir que todas las medias son significativamente distintas, siendo el método de inoculación 3 el que presenta un mayor peso seco medio.

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	valor P	(Error)
Modelo	351.11	11	31.92	55.03	<0.0001	
BLOQUE	27.85	3	9.28	2.78	0.2120	
CULTIVAR	187.60	1	187.60	56.10	0.0049	(BLOQUE*CULTIVAR)
BLOQUE*CULTIVAR (Error(a))	10.03	3	3.34	5.77	0.0112	
INOC	122.76	2	61.38	105.83	<0.0001	
CULTIVAR*INOC	2.86	2	1.43	2.47	0.1265	
Error (Error(b))	6.96	12	0.58			
Total	358.07	23				

Test:Tukey Alfa:=0.05 DMS:=1.01591

Error: 0.5800 gl: 12

INOC	Medias	n			
1.00	30.34	8	A		
2.00	32.51	8		B	
3.00	35.84	8			C

Letras distintas indican diferencias significativas(p<= 0.05)

FACTORES ALEATORIOS Y MODELOS MIXTOS

El estudio detallado de los modelos mixtos supera el nivel de este curso, en consecuencia sólo presentaremos una breve introducción conceptual que debería servir para facilitar al lector: 1) el abordaje del tema en textos más avanzados y 2) la posibilidad de reconocer si un problema particular corresponde a este capítulo de la estadística. El nombre modelos *mixtos* alude a que los elementos del modelo son una mezcla de lo que se denomina *factores fijos* y *factores aleatorios*. A continuación explicaremos el significado de estos términos y se pondrá en evidencia que todos los factores y tratamientos estudiados en los experimentos que fuimos presentando en el curso hasta el momento, corresponden a la primera categoría, la de factores fijos.

Los ***factores fijos*** son variables explicativas para las cuales los niveles que se ponen a prueba en el experimento fueron específicamente seleccionadas por el investigador. Cada nivel de interés resultó incluido en el experimento. No hay interés en otros niveles y las pruebas de comparaciones múltiples se emplean para determinar cuáles son los pares de medias que

medias que difieren entre sí. La hipótesis nula referida a un factor fijo postula que la variable respuesta de interés no difiere en su respuesta a los distintos niveles del factor en cuestión. Ejemplos de factores fijos pueden ser especies, dietas, variedades, densidades de siembra, clases de fertilizantes, tipo de labranza, para los cuales el investigador esta interesado en poner a prueba la hipótesis nula que dice que los efectos de las distintas dietas, variedades, etc., son iguales. Estos factores se denominan ***fijos*** porque en caso de repetirse el experimento se volverían a utilizar los mismos niveles del factor.

Los ***factores aleatorios*** son variables explicativas para las cuales el investigador no elige específicamente determinados niveles a ser incorporados en su ensayo, sino que los selecciona al azar a partir de una población mayor de niveles posibles. En consecuencia, desea extraer conclusiones acerca de la población entera de donde los extrajo. En la mayoría de los casos está interesado en obtener una estimación de la magnitud de la componente de varianza, es decir la variabilidad debida a la incorporación de ese factor en el modelo. A diferencia de los factores fijos, no hay comparaciones entre poblaciones sino que hay una sola población y el interés está centrado en estimar cuánto de la variación en la respuesta se puede adjudicar a ese factor. La hipótesis nula referida a un factor aleatorio postula que la componente de la varianza en la variable respuesta asociada a los distintos niveles de ese factor es nula. Ejemplos de factores aleatorios pueden ser genotipos, individuos, familias, empresas agropecuarias, localidades. En caso de repetirse el experimento, el investigador seleccionaría una nueva muestra de niveles del factor (otros genotipos, otros individuos, otras familias, otras empresas u otras localidades).

Estas características condensan los tres criterios principales para decidir si un factor debe considerarse fijo o aleatorio:

1. ¿Los niveles del factor se seleccionaron porque son de particular interés o se eligieron al azar?
2. ¿Las conclusiones se van a confinar a los niveles del factor que fueron estudiados o se van a aplicar a una población mayor de niveles? Por ejemplo si concluiremos que hay diferencias en la adopción de tecnología entre las familias Pérez, Nakandare y Osakia o que hay variabilidad en la adopción de tecnología entre familias de floricultores de Florencio Varela (de las cuales aquellas constituyen una muestra que participó en el estudio).
3. Si el experimento se repitiera, ¿se estudiarían nuevamente los mismos niveles del factor o se tomaría una nueva muestra a partir de la población mayor de niveles?

Así como difieren en el tipo de inferencias acerca de los parámetros poblacionales, los factores aleatorios también difieren en las esperanzas de los cuadrados medios y en consecuencia, en los denominadores adecuados para construir los cocientes F de las pruebas. La interpretación de las interacciones también es diferente entre los modelos mixtos y los modelos de factores fijos. Consideremos un experimento con factor A fijo y B aleatorio, para una determinada variable respuesta, la hipótesis nula referida a A postula que no hay diferencias entre promedios de los niveles de A, para el factor B la hipótesis nula postula que no hay variabilidad entre los niveles de B (incluyendo aquellos que no participan del ensayo). Para el término de interacción (Ax B) la hipótesis nula postula que la variabilidad entre niveles de B es la misma a través de los niveles de A. En el caso de dos factores fijos (A y

C por ejemplo), la hipótesis nula postula que la respuesta a los distintos niveles de A no difiere entre niveles específicos de C.

Por ejemplo, si se desea estudiar la variabilidad de la respuesta a la inundación de una población de *Paspalum dilatatum* (pasto miel, una gramínea forrajera nativa), el factor A: disponibilidad hídrica, podría tener tres niveles: sin inundación, inundación de corta duración (una semana) e inundación prolongada (un mes), seleccionados por el investigador como únicos niveles de interés (las conclusiones del estudio se referirán sólo a estos tres regímenes de agua en el suelo), mientras el factor B: clones (grupos de individuos con igual información genética) podría tener 8 niveles. Entonces el investigador debe seleccionar al azar 8 plantas de la población de pasto miel y multiplicarlas en el invernáculo para tener repeticiones genéticamente idénticas, entre las cuales asignará al azar a un tercio de cada grupo cada tratamiento de inundación. Supongamos que toma como variable respuesta el peso de la planta al finalizar la estación de crecimiento y veamos cuáles son las hipótesis de interés.

Para el factor A: inundación (fijo), la hipótesis nula postula que no hay diferencias en el peso promedio de las plantas entre los tres niveles de inundación ensayados. Para el factor B: clones (aleatorio), la hipótesis postula que no hay variabilidad en el peso de las plantas de la población estudiada (no entre las 8 plantas particulares que el investigador extrajo sino entre las plantas de la población de la cual estas 8 fueron seleccionadas). Para la interacción A x B , la hipótesis postula que la respuesta a los tratamientos de inundación no varía entre las plantas de la población. Al igual que en los experimentos factoriales de factores fijos, primero se pone a prueba la presencia de interacción y si no se detecta interacción significativa se procede al estudio de los efectos principales de los factores.

LECTURAS RECOMENDADAS

Kuehl, R. 2001. Diseño de experimentos. Segunda Edición. Thomson. Capítulos 13 y 14.

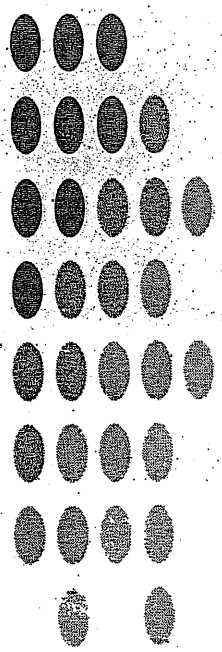
Steel, R. G. y J. H. Torrie. 1985. Bioestadística: Principios y Procedimientos. Segunda edición. McGraw-Hill. Capítulos 15 y 16.

Capítulo 7
Análisis de regresión

El estudio de la relación funcional entre variables cuantitativas.

Predicción de una respuesta mediante varias variables explicativas.

Los residuales como elemento de diagnóstico



ANÁLISIS DE REGRESIÓN

En términos generales se puede decir que el análisis de regresión se refiere al estudio de la dependencia de una variable, la variable respuesta o dependiente, respecto a una o más variables adicionales, las variables explicativas, con la perspectiva de estimar y/o predecir el valor medio o promedio poblacional de la primera para valores conocidos o fijos de las segundas. Por ejemplo:

- Un agrónomo, puede estar interesado en estudiar la dependencia existente entre los niveles de producción de un cultivo, por ejemplo de trigo, y la temperatura, la lluvia, la cantidad de luz solar y el nivel de fertilidad del suelo. Este análisis de dependencia puede permitirle predecir o pronosticar la producción promedio de ese cultivo, dada la información existente sobre las variables explicativas.

- El director de marketing de una empresa puede estar interesado en conocer la manera como se relacionan la demanda de su producto con los gastos en publicidad en que incurre dicha empresa. Este tipo de estudio sería de gran utilidad para averiguar la elasticidad de la demanda del producto en los gastos de publicidad de la empresa, es decir, la respuesta promedio de la demanda ante un aumento de una unidad en el presupuesto de gastos en publicidad. Este conocimiento a su vez puede ser de mucha utilidad para determinar el presupuesto "óptimo" de la empresa.

El propósito principal de estos estudios es obtener una función sencilla de la/s variable/s explicativa/s, que sea capaz de describir lo más ajustadamente posible la variación de la variable respuesta. Como los valores observados de la variable respuesta generalmente difieren de los que predice la función, ésta posee un error. La función más eficaz es aquella que describe la variable dependiente con el menor error posible o, dicho en otras palabras, con la menor diferencia entre los valores observados y los valores predichos por el modelo. Estas diferencias entre los valores observados y los predichos (el error de la función) se denominan residuales.

El análisis de regresión se propone:

Estimar la relación funcional entre las **x** y la **y**

Formular hipótesis respecto de los parámetros del modelo

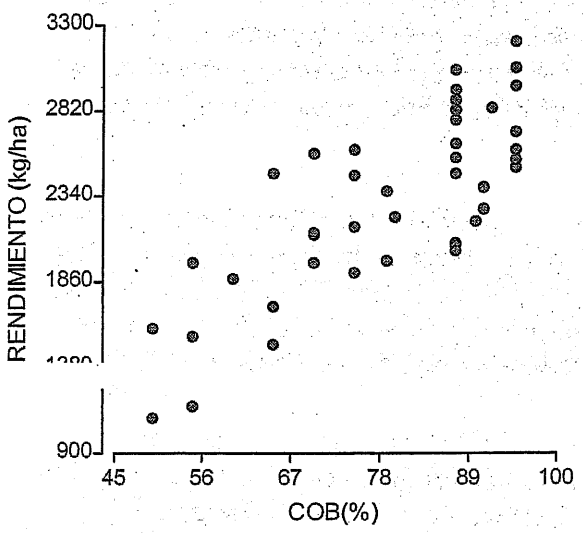
Predecir valores de la variable respuesta para valores específicos de las variables explicativas.

Para estimar los parámetros de la función se utiliza el ajuste por mínimos cuadrados. Es decir, se trata de encontrar la función en la cual la suma de los cuadrados de las diferencias entre los valores observados y los que predice la función estimada sea mínima.

7.1. Regresión lineal con una variable independiente

El siguiente diagrama de dispersión representa los rendimientos de 42 lotes cultivados con girasol en partidos del oeste de la Provincia de Buenos Aires, representados en fun-

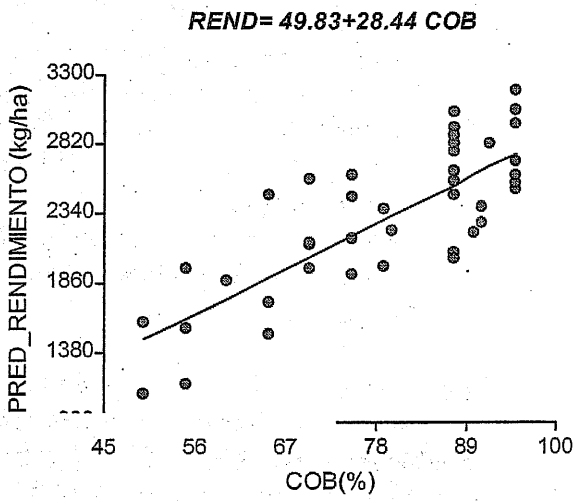
ción de la cobertura del suelo generada por el cultivo al momento de la floración en el mismo lote.



La figura sugiere que hay una relación lineal entre la cobertura en el momento de la floración y los rendimientos, y muestra que aquellos lotes que lograron mayores coberturas al momento de floración, alcanzaron mayores rendimientos. Sin embargo, la relación no es perfecta ya que los puntos no se encuentran exactamente alineados.

La siguiente figura muestra la recta que describe la relación estadística entre las variables estudiadas.

La dispersión de puntos alrededor de la línea representa la variación en los rendimientos no explicada por los niveles de cobertura del cultivo en floración.



7.1.1. Modelo de Regresión

El modelo de regresión lineal con una sola variable independiente es:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

donde:

y_i es el valor de la variable respuesta en la observación i

β_0 y β_1 son parámetros ordenada al origen y pendiente, respectivamente

x_i es una constante conocida, el valor de la variable independiente en la observación i

ε_i es un término error aleatorio con distribución normal, media $E(\varepsilon_i) = 0$ y variancia $\sigma^2(\varepsilon_i) = \sigma^2$ (por ende, la variancia del error es igual para todos los niveles de X), ε_i y ε_j no están correlacionados de modo que la covarianza $\sigma(\varepsilon_i, \varepsilon_j) = 0$ para todo i, j; con $i \neq j$.

Este modelo se dice que es simple, lineal en los parámetros, y lineal en la variable independiente. Es "simple" porque sólo hay una variable independiente, es "lineal en los parámetros" porque los parámetros no aparecen como exponentes o multiplicados o divididos por otro parámetro, y es "lineal en la variable independiente" porque la variable independiente aparece en la primera potencia.

7.1.2. Consideraciones importantes del modelo

- El valor observado de **Y** en la observación i es la suma de dos componentes:

  Una componente constante $\beta_0 + \beta_1 X_i$

  Un término aleatorio ε_i,.

  Por lo que se desprende que $\mathbf{Y_i}$ es una variable aleatoria.

- Como $E(\varepsilon_i) = 0$, entonces

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i + E(\varepsilon_i) = \beta_0 + \beta_1 X_i$$

  Por lo tanto, hay una distribución de probabilidad de **Y** para cada nivel de X y las medias de estas distribuciones varían en forma sistemática con **X**.

  Así, la respuesta $\mathbf{Y_i}$, cuando el nivel de **X** en la observación i es X_i , proviene de una distribución de probabilidad cuya media es:

$$E(Y_i) = \beta_0 + \beta_1 X_i$$

- El valor observado de **Y** en la observación i excede o cae por debajo del valor de la función de regresión por la cantidad dada por el término error, ε_i.
- Se asume que los términos error ε_i tienen varianza constante, σ^2. Se desprende, de aquí, que la varianza de la respuesta $\mathbf{Y_i}$ es:

$$\sigma^2_{(Yi)} = \sigma^2$$

  ya que $\sigma^2_{(\beta 0 + \beta 1 Xi + \varepsilon i)} = \sigma^2_{(\varepsilon i)} = \sigma^2$

  Así, el modelo asume que las distribuciones de probabilidad de Y tienen la misma varianza σ^2, sin importar el nivel de la variable X.

- Se asume que los términos error no están correlacionados. Por lo tanto, el resultado en un ensayo u observación no tiene efecto en el término error de otro ensayo u observación (en cuanto a si es grande o pequeño, negativo o positivo). Así como los términos error no están correlacionados, las respuestas Y_i y Y_j tampoco lo están.

En síntesis, el modelo tiene como supuestos que las observaciones de la variable respuesta Y_i provienen de distribuciones de probabilidad cuyas medias son $E(Y_i) = \beta_0 + \beta_1 X_i$ y cuyas varianzas σ^2, son las mismas para todos los niveles de X. Además cualesquiera dos observaciones Y_i y Y_j no están correlacionadas entre sí.

7.1.3. Estimación de la función de regresión

Los parámetros β_0 y β_1 se denominan coeficientes de regresión, β_0 es la ordenada al origen de la recta. Si el alcance del modelo incluye x=0, β_0 es la media de la distribución de probabilidad de **Y** cuando **X**=0. β_1 es la pendiente de la recta de regresión, indica el cambio en la media de la distribución de probabilidad de **Y** por cada aumento unitario de **X**.

Debemos estimar β_0 y β_1 de los datos de la muestra. Para encontrar "buenos" estimadores (insesgados y de mínima varianza) de estos parámetros, se utiliza el método de mí-

nimos cuadrados. El método parte de la suma de las n desviaciones cuadráticas de los valores observados respecto a los predichos. Si denominamos Q a la expresión:

$$Q= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

los estimadores de β_0 y β_1 son aquellos valores de $\hat{\beta}_0$ y $\hat{\beta}_1$ respectivamente, que minimizan Q para una muestra dada de observaciones.

Los **estimadores mínimos cuadráticos de los parámetros de la regresión** son los siguientes:

$$\hat{\beta}_1 = \frac{SP_{xy}}{SC_x} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad \hat{\sigma}^2 = S^2 = \frac{SC_E}{n-2}$$

donde:

$$SC_x = \sum_{i=1}^{n}(x_i - \bar{x})^2 , \qquad SP_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) , \qquad SC_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Una vez que han sido obtenidos los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ de los parámetros de la función de regresión, es natural que estimemos la función de regresión como:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

donde $\hat{Y}_i$ es la respuesta media estimada correspondiente al nivel X_i de la variable independiente.

7.1.4. Residuales

La diferencia entre un valor observado cualquiera de Y y el valor ajustado correspondiente $\hat{Y}$ se llama residual (e). Así, el residual para la observación i se puede escribir: $e_i = Y_i - \hat{Y}_i$

7.1.5. Inferencia acerca de la pendiente β1

Frecuentemente la primera inferencia que se desea hacer al estudiar la relación entre **Y** y **X** se refiere a la existencia misma de la relación. ¿Aporta **X** información para predecir **Y**? Es decir, ¿presentan los datos información suficiente que indique que **Y** aumenta (o disminuye) linealmente cuando **X** crece en el rango de valores observados?

La pregunta práctica que planteamos se refiere al valor de β_1, que es el cambio medio en Y por unidad de cambio de X. Afirmar que Y no aumenta (o disminuye) linealmente cuando X aumenta equivale a decir que $\beta_1 = 0$. Entonces se prueba la hipótesis nula que postula que $\beta_1 = 0$ contra la alternativa $\beta_1 \neq 0$.

Asumiendo que el error aleatorio ε tiene una distribución normal, se puede demostrar que la distribución muestral de $\hat{\beta}_1$ tendrá una distribución normal y que el valor esperado y la variancia de $\hat{\beta}_1$ serán:

$$E(\hat{\beta}_1) = \beta_1 \qquad \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{SC_x}$$

El estimador de la variancia σ^2 es S^2, y el estadístico de prueba adecuado es

$t^* = \dfrac{\hat{\beta}_1 - \beta_1}{S / \sqrt{SC_x}}$ que sigue una distribución t de Student con n-2 grados de libertad.

Si $|t^*|$, calculado a partir de los datos de la muestra es mayor que el valor de tabla $t_{(1-\alpha/2;\, n-2)}$ se rechaza la hipótesis de que no existe regresión lineal.

Una vez que se ha determinado que X aporta información para la predicción de Y, interesa conocer cuál es el cambio pronosticado en Y por unidad de cambio en X y cuánta confianza puede tenerse en la estimación.

Puede demostrarse que el intervalo de confianza para β_1 con un coeficiente de confianza $(1-\alpha)$ es:

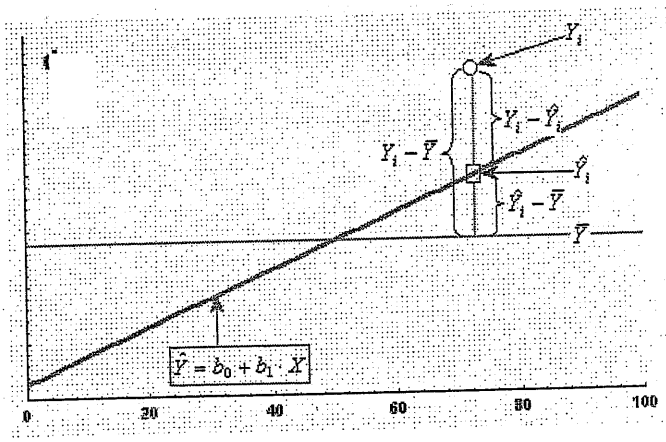
$$\hat{\beta}_1 \pm t_{(n-2;1-\alpha/2)} \frac{S}{\sqrt{SC_x}}$$

Notar que tanto para la construcción de intervalos de confianza como para poner a prueba hipótesis acerca de los parámetros de la regresión se agrega a los supuestos ya enunciados anteriormente para el modelo, el supuesto de normalidad, que no era necesario para la estimación de parámetros mediante el método de mínimos cuadrados.

7.1.6. ANOVA para el análisis de regresión

Si repartimos la variación que aporta cada una de las observaciones al total (su desvío con respecto a la media), como se observa en la figura, en distintas fuentes de variación, obtenemos:

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$



La primera diferencia del segundo miembro de la igualdad corresponde el valor del desvío del valor de la recta ajustada alrededor de la media. Es la desviación explicada ya que explica en cuanto se reduce la desviación total cuando la recta de regresión se ajusta a los puntos. La segunda diferencia corresponde al desvío de una observación alrededor de la recta de regresión, es la desviación no explicada relacionada con el error aleatorio.

Elevando al cuadrado y sumando para todos los valores obtenemos:

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$$

Es decir que:

$SC_{TOTAL} = SC_{REGRESIÒN} + SC_{ERROR}$

La SC_R tiene 1 grado de libertad asociado pues hay una sola variable independiente en el modelo, los grados de libertad de la SC_E son n –2 pues fueron estimados 2 parámetros y la SC_{TOTAL} tiene n-1 grados de libertad asociados. Entonces: n -1 = 1 + (n -2)

Se construyen los correspondientes cuadrados medios (CM)

$CM_R = SC_R / 1$ $\qquad\qquad$ $CM_E = SC_E / (n-2)$

Donde CM_E es el ***cuadrado medio del error***:

$$CM_E = \frac{SC_E}{n-2} = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum(Y_i - b_0 - b_1 \cdot X_i)^2}{n-2} = \frac{\sum e_i^2}{n-2}$$

Puede demostrarse que CM_E es un estimador insesgado de σ^2 para el modelo de regresión: $E(CM_E) = \sigma^2$. En consecuencia, un estimador de la desviación standard σ, es simplemente la raíz cuadrada positiva de CM_E.

También se puede demostrar que:

$$E(CM_R) = \sigma^2 + \beta_1^2 \sum (X_i - \overline{X})^2$$

Entonces, en relación con las siguientes hipótesis nula y alternativa:

Ho: $\beta_1 = 0$ $\qquad\qquad$ H1: $\beta_1 \neq 0$

Si vale H_0 el cociente $E(CM_R) / E(CM_E) = 1$

En cambio, si vale H_1 este cociente tiende a ser más grande que 1, pero como no se puede trabajar con los valores esperados pues son parámetros desconocidos, se usan sus estimadores. Así si CM_R / CM_E da un valor grande, sugerirá que $\beta_1 \neq 0$.

Se prueba que bajo el supuesto de normalidad el cociente $CM_R/CM_E = F^*$ tiene distribución $F_{1,n-2}$ bajo H_0. Dado que la prueba es unilateral derecha (rechaza para valores grandes del cociente), la regla de decisión, con una probabilidad de *error de tipo I* igual a α, es no rechazar H_0 si $F^* \leq F_{1-\alpha;1;n-2}$ y rechazar H_0 si $F^* > F_{1-\alpha;1;n-2}$. Esta información se resume en una tabla de ANOVA como la siguiente:

F de Variación	SC	gl	CM	F
Regresión	SC_R	1	CM_R	CM_R / CM_E
Error	SC_E	n - 2	CM_E	
Total	SC_{TOTAL}	n - 1		

Para el ejemplo de los rendimientos de 42 lotes de girasol presentado al comienzo del capítulo se muestran a continuación los coeficientes estimados, sus errores estándar y los límites de los intervalos de confianza para los parámetros β_1 y β_0.

Coeficientes de regresión y estadísticos asociados

Coef	Estimación	Err.Est.	LI(95%)	LS(95%)	t	valor p
const	49.83	285.45	-527.1	626.75	0.17	0.8623
COB(%)	28.44	3.55	21.27	35.61	8.02	<0,0001

La tabla de Análisis de Varianza para el mismo ejemplo es la siguiente:

F.V.	SC	gl	CM	F	valor p
COB(%)	6399033	1	6399032.97	64.27	<0,0001
Error	3982862.7	40	99571.57		
Total	10381896	41			

Como se puede observar, el coeficiente asociado a la cobertura (β_1) es significativamente distinto de cero ($p<0.0001$). Su valor estimado indica que, dentro del rango de coberturas abarcado por el estudio, por cada aumento de 1% en la cobertura en el momento de la floración, el rendimiento final aumenta en promedio 28.44 kg/ha. Ciertamente, en la regresión simple el ANOVA no agrega información adicional a la que proveen la prueba t de la pendiente y el cálculo del coeficiente de determinación. Sin embargo, su utilidad se pondrá en evidencia cuando abordemos modelos más complejos.

7.1.7. Estimación del valor esperado de Y para un valor dado de X

La estimación del valor medio de Y para un valor dado de X [es decir, la estimación de E(Y/X)] puede ser de interés práctico. Si el rendimiento del cultivo de girasol (Y) está relacionado linealmente con la cobertura al momento de floración (X), el productor puede estar interesado en estimar el rendimiento promedio esperado para lotes que presentan una determinada cobertura.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

entonces se utiliza $\hat{Y}$ para estimar el valor esperado de Y para un dado valor de X, digamos x_p. Puede demostrarse que $\hat{Y}$ tiene distribución normal con varianza

$$\sigma^2_{\hat{y}} = \sigma^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}\right]$$

La varianza estimada de $\hat{Y}$ emplea S^2 en lugar de σ^2 en la expresión anterior. El intervalo de confianza para el valor esperado, con un coeficiente de confianza de $(1-\alpha)$ es

$$\hat{Y} \pm t_{(n-2;\alpha/2)} S\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SC_x}}$$

Para el ejemplo de los lotes cultivados con girasol se obtuvo la ecuación estimada:

$$\hat{Y} = 49.83 + 28.44\ X$$

Para una cobertura de 60% se estima un rendimiento medio de 1756.23 kg/ha. Teniendo en cuenta que $S^2 = CM_E = 99571.57$, que la $SC_x = 7909.72$ y la media de las coberturas es 79.27; se puede estimar un intervalo para el rendimiento promedio esperado en lotes con 60% de cobertura:

$$1756.23 \pm 2.328 \cdot 315.54 \cdot \sqrt{\frac{1}{42} + \frac{(60-79.27)^2}{7909.72}}$$, que resulta: 1756.23 ± 194.71

es decir que el rendimiento promedio esperado estará cubierto por el intervalo 1561.52kg/ha a 1950.94kg/ha con una confianza del 95%.

7.1.8. Predicción de un valor particular de Y para un valor dado de X

Consideraremos ahora la predicción de una nueva observación Y correspondiente a un dado nivel X de la variable independiente. Por ejemplo, en el caso de los cultivos de

girasol, se podría querer predecir el rendimiento para un lote particular conociendo la cobertura al momento de la floración.

Se considera a la nueva observación que se va a estimar, como el resultado de un nuevo ensayo, independiente de los ensayos sobre los cuales se ha basado el análisis de regresión. Denotaremos al nivel de X para el nuevo ensayo como X_h y a la nueva observación Y como $Y_{h(n)}$. Por supuesto que asumiremos que el modelo de regresión subyacente aplicable a los datos básicos de la muestra sigue siendo apropiado para la nueva observación.

La diferencia entre la estimación de la respuesta media, $E(Y_h)$ - discutida en la sección anterior y la predicción de una nueva respuesta, $Y_{h(n)}$, que estamos discutiendo ahora, es importante. En el primer caso, estamos estimando la **media** de la distribución de Y. Aquí, en cambio, estamos prediciendo un **resultado individual** extraído de la distribución de Y. Obviamente, la gran mayoría de los resultados individuales se desvían de la respuesta media, y esto debe ser permitido por el procedimiento de predicción de $Y_{h(n)}$.

Los límites de predicción para una nueva observación Y a un dado nivel de X_h se obtienen por medio del siguiente teorema:

$$\frac{\hat{Y}_h - Y}{s\left(Y_{h(n)}\right)} \text{ tiene distribución } t_{n-2}$$

Nótese que este estadístico estandarizado utiliza al estimador puntual $\hat{Y}_h$ en el numerador y no la verdadera media $E(Y_h)$ porque la verdadera media es desconocida y no puede ser utilizada para hacer una predicción. Definiremos luego la desviación standard estimada, $s(Y_{h(n)})$ que aparece en el denominador del estadístico. A partir de este último teorema se deduce que los límites de predicción de 1 - α para una nueva observación son:

$$\hat{Y}_h \pm t_{1-\alpha/2;n-2} \cdot s\left(Y_{h(n)}\right)$$

Denotaremos a la varianza del denominador del estadístico como $\sigma^2(Y_{h(n)})$ y es igual a:

$$\sigma^2\left(Y_{h(n)}\right) = \sigma^2\left(\hat{Y}_h - Y\right) = \sigma^2\left(\hat{Y}_h\right) + \sigma^2 .$$

Notemos que $\sigma^2(Y_{h(n)})$ tiene dos componentes:

1) la varianza de la distribución por muestreo de $\hat{Y}_h$ y,
2) la varianza de la distribución de Y.

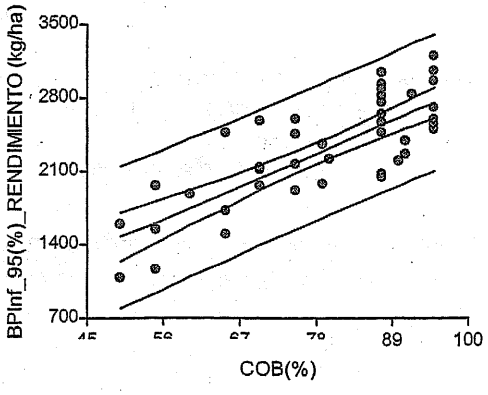
Un estimador insesgado de $\sigma^2(Y_{h(n)})$ es $s^2\left(Y_{h(n)}\right) = s^2\left(\hat{Y}_h\right) + CM_E$

lo cual puede expresarse como:

$$s^2\left(Y_{h(n)}\right) = CM_E \cdot \left[1 + \frac{1}{n} + \frac{\left(X_h - \overline{X}\right)^2}{\sum\left(X_i - \overline{X}\right)^2}\right]$$

El gráfico muestra las bandas de confianza y de predicción para el ejemplo de los lotes cultivados con girasol.

Bandas de confianza y de predicción



7.1.9. Coeficiente de determinación R2

El coeficiente de determinación es una medida descriptiva del grado de asociación lineal entre las dos variables y al mismo tiempo, de la bondad de ajuste del modelo. La SC_{TOTAL} mide la variación total en las observaciones Y_i mientras que la SC_E mide la variación residual en las Y_i cuando se emplea el modelo de regresión que usa a la variable X. Una medida natural del efecto de X para reducir la variación en Y es:

$$R^2 = \frac{SC_{TOTAL} - SC_E}{SC_{TOTAL}} = \frac{SC_R}{SC_{TOTAL}}$$

Dado que $0 \leq SC_R \leq SC_{TOTAL}$, entonces $0 \leq R^2 \leq 1$.

Este coeficiente indica cuánto de la variación total de Y puede ser explicado por la dependencia lineal de X. En el ejemplo de los lotes cultivados con girasol $R^2 = SC_R/SC_{TOTAL} = 6399033/\ 10381896 = 0.616$, es decir que la cobertura en floración explica aproximadamente el 62% de la variación de los rendimientos de los lotes. Si no existe regresión lineal, $R^2 = 0$, $SC_E = SC_{TOTAL}$, lo que indica que $SC_R = 0$. No hay asociación lineal entre X e Y y la variación en X no es de ninguna ayuda para explicar la variación de las observaciones Y_i. En la práctica no es probable que R^2 sea 0 o 1, sino que se encuentra entre ambos valores. Cuanto más cercano a 1 sea el valor, más grande es el grado de asociación lineal entre X e Y.

7.2. REGRESIÓN MÚLTIPLE

El análisis de regresión múltiple es una extensión del análisis de regresión simple para la situación en que se consideran dos o más variables explicativas en el modelo. En esta parte del capítulo se presentarán algunos tipos particulares de modelos de regresión múltiple, los supuestos relacionados con el modelo general, las pruebas de hipótesis referidas a los parámetros principales y las conclusiones derivadas del análisis. Antes de introducir estos temas, es importante puntualizar que tratar con varias variables independientes simultáneamente es considerablemente más dificultoso que el análisis de regresión simple, por las siguientes razones:

1. Es más difícil determinar la mejor elección del modelo cuando hay varias variables candidatas a explicar la variación de la respuesta.
2. Es más difícil visualizar gráficamente el modelo ajustado y las observaciones (especialmente en el caso de más de dos variables explicativas ya que es imposible realizar figuras de más de tres dimensiones).
3. En algunas circunstancias es difícil interpretar biológicamente el significado del modelo de mejor ajuste.
4. Los cálculos del análisis requieren conocimientos de álgebra matricial y prácticamente no se realizan a mano, sino con paquetes estadísticos en computadoras.

En parte debido a estas razones, la regresión múltiple conforma un campo mucho más amplio del que este curso puede abarcar. Por ejemplo, existen distintos métodos para seleccionar las mejores variables explicativas así como para comparar diferentes modelos, aspectos que no son cubiertos en este curso. En
las siguientes páginas se presentará primero el modelo más sencillo de regresión múltiple, que consta de dos variables explicativas independientes entre sí. Luego se mostrará el modelo general de regresión múltiple, se analizará un ejemplo y se estudiarán dos

tipos particulares de regresión múltiple: la regresión polinómica y la regresión con variables categóricas.

7.2.1 Modelo de primer orden con dos variables independientes

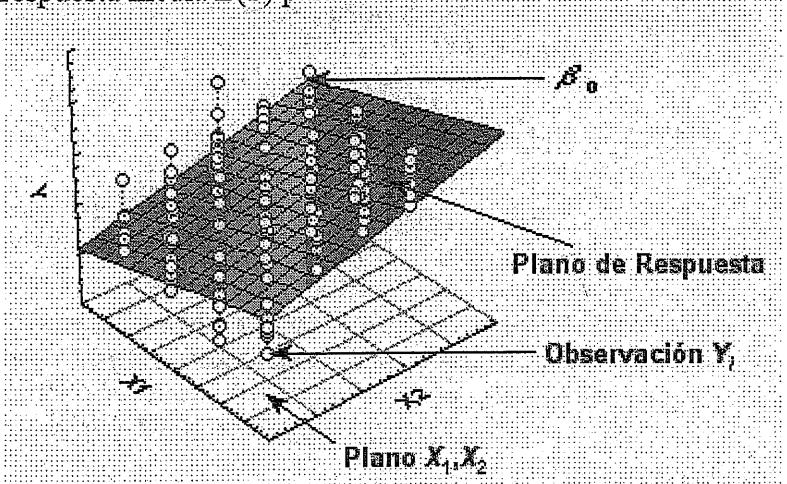
Cuando hay dos variables independientes X_1 y X_2, el modelo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \qquad [1]$$

es un modelo lineal en los parámetros (β_0, β_1 y β_2) y en las variables independientes. Los ε_i son los términos error, donde ε_i son independientes y se distribuyen normalmente con parámetros $\mu = 0$ y variancia σ^2. Entonces, asumiendo que $E(\varepsilon_i) = 0$, la función de regresión del modelo [1] es:

$$E(Y_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \qquad [2]$$

Geométricamente, la función de regresión [2] es un **plano tridimensional de respuesta**. La siguiente figura contiene una representación de una porción del plano de respuesta: $E(Y_i) = 15.48 + 0.06 X_1 - 1.23 X_2$. Un punto que está sobre el plano de respuesta corresponde a la respuesta media $E(Y)$ para una dada combinación de niveles de X_1 y X_2.



La figura también muestra una serie de observaciones Y_i correspondientes a los distintos niveles de las variables independientes (X_{i1} y X_{i2}). La distancia vertical desde Y_i hasta el plano de respuesta representa el error $\varepsilon_i = Y_i - E(Y_i)$. A menudo la función de regresión se denomina **superficie de respuesta.**

7.2.2. Significado de los Coeficientes de Regresión

El parámetro β_0 es la intercepción del plano de regresión con el eje Y. Si el alcance del modelo incluye $X_1 = 0$ y $X_2 = 0$, β_0 da la respuesta media en $X_1 = 0$ y $X_2 = 0$. El parámetro β_1 indica el cambio en la respuesta media por cada unidad de incremento en X_1 cuando X_2 se mantiene constante. De la misma manera, β_2 indica el cambio en la respuesta media por cada unidad de incremento en X_2 cuando X_1 se mantiene constante. Para ver esto en nuestro ejemplo, supongamos que X_2 se mantiene constante al nivel $X_2 = 40$. La función de regresión es ahora: $E(Y) = 15.48 + 0.06 X_1 - 1.23 (40) = -33.72 + 0.06 X_1$.

Nótese que para $X_2 = 40$, la función de respuesta es una línea recta con pendiente igual a 0.06. Lo mismo es cierto para cualquier otro valor de X_2; solo la intercepción de la función de respuesta cambia. Por tanto, $\beta_1 = 0.06$ indica que la respuesta media se incrementa en 0.06 con un incremento de una unidad en X_1 cuando X_2 se mantiene constante, sin importar cuál sea el nivel de X_2. Similarmente, $\beta_2 = -1.23$ indica que la respuesta media decrece en 1.23 unidades con un incremento unitario en X_2 cuando X_1 se mantiene constante. Cuando el efecto de X_1 sobre la respuesta media no depende del nivel de X_2 y, correspondientemente, el efecto de X_2 sobre la respuesta media no depende del nivel de X_1, se dice que las dos variables independientes tienen **efectos aditivos** o que no presentan **interacción**.

Los parámetros β_1 y β_2 se denominan **coeficientes de regresión parcial** porque reflejan el efecto parcial de una variable independiente cuando la otra variable independiente es incluida en el modelo y es mantenida constante.

Se puede calcular una región de confianza para toda la superficie de regresión. Esta es una extensión del procedimiento para construir una banda de confianza para la recta, que fue presentado para la regresión simple.

Se debe evitar hacer extrapolaciones fuera del conjunto observado de valores de las variables X1y X2. No es suficiente determinar la amplitud de cada una por separado para definir esa región. Por ejemplo, graficados en dos dimensiones los X1y X2 observados, pueden estar comprendidos en una elipse de menor superficie que el rectángulo determinado por la amplitud de cada una de las variables.

7.2.3. Modelo general de Regresión Lineal Múltiple

Consideremos ahora el caso donde hay $p - 1$ variables independientes $X_1, X_2,...,X_{p-1}$. El modelo:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

donde:

$\beta_0, \beta_1, \ldots, \beta_{p-1}$ son parámetros;
$X_{i1}, X_{i2},...,X_{i,p-1}$ son constantes conocidas ;
ε_i son variables aleatorias independientes $N(0,\sigma^2)$, $i = 1,2,...,n$.

La función de respuesta es:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1}.$$

Esta función de respuesta es un **hiperplano**, que es un plano en más de dos dimensiones. Ya no es posible representar gráficamente esta superficie de respuesta como lo hicimos en la Figura 1 para el caso de dos variables independientes, pero el significado de los parámetros es análogo. El parámetro β_k indica el cambio en la respuesta media $E(Y)$ frente a una unidad de incremento en la variable independiente X_k cuando todas las otras variables independientes incluidas en el modelo se mantienen constantes.

Cuando $X_1,...,X_{p-1}$ representan diferentes variables independientes, el modelo lineal general es un modelo de primer orden en el cual no hay efectos de interacción entre las variables independientes.

La variación explicada por el modelo se puede a su vez particionar para estudiar la contribución relativa de las distintas variables explicativas. Generalmente es de interés estudiar tanto la contribución parcial de cada una cuando todas las demás variables ya participan en el modelo, como la explicación secuencial de la variación total que aporta cada variable al entrar al modelo en la secuencia propuesta. En el primer caso las sumas de cuadrados calculadas se denominan ***SC parciales*** y en el segundo ***SC secuenciales***. El significado de las distintas SC quedará más claro en el análisis del ejemplo que se presenta más abajo. Allí también podremos comprobar que sólo las SC secuenciales constituyen una verdadera partición de la SC del modelo (SC_R).

7.2.7. Coeficiente de determinación múltiple

$$R^2 = \frac{SC_{TOTAL} - SC_E}{SC_{TOTAL}} = \frac{SC_R}{SC_{TOTAL}}$$

El R^2 varía entre 0 y 1. $R^2 = 0$ si todos los $\beta_k = 0$, con $1 \leq k \leq p-1$, en cambio $R^2 = 1$ si todas las observaciones caen sobre la superficie respuesta ajustada. El R^2 representa la proporción de la variación en la variable Y que ha sido explicada por las variables independientes X_1,....., X_{p-1}, de acuerdo al modelo propuesto-

7.2.8. Ejemplo de regresión lineal múltiple

En un experimento para evaluar el efecto de algunas variables sobre la erosión del suelo, se aplicó una lluvia artificial durante 20 minutos a 11 parcelas cuadradas de 50 cm de lado, ubicadas en un paisaje ondulado. En cada parcela se registró la pérdida de suelo en g/m^2 (PERDS: variable respuesta), el gradiente topográfico medido como porcentaje de inclinación (INCL: X1), la proporción de suelo cubierto por la vegetación (COBVEG: X2) y el diámetro de la mayor abertura de suelo desnudo medido en cm (ABMAX: X3).

Parcela	1	2	3	4	5	6	7	8	9	10	11
PERDS(Y)	2,71	3,56	3,14	3,78	4,02	3,98	5,55	4,36	5,21	4,38	3,57
INCL(X1)	0,43	0,47	0,44	0,48	0,48	0,49	0,53	0,5	0,55	0,51	0,48
COBVEG(X2)	0,34	0,32	0,29	0,3	0,25	0,26	0,1	0,16	0,19	0,18	0,28
ABMAX(X3)	1,95	5,13	3,98	6,25	7,12	6,5	10,67	7,08	9,88	8,72	4,96

Por el momento olvidaremos que X3 fue registrada y consideraremos un modelo con dos variables regresoras (X1 y X2). A continuación se presentan los estimadores de los parámetros, sus respectivos errores estándar, intervalos de confianza y pruebas *t* de significación.

Coeficientes de regresión y estadísticos asociados

Coeficiente	Estimador	EE	LI(95%)	LS(95%)	T	valor p
constante	-2,67	1,66	-6,5	1,16	-1,61	0,1468
INCL(X1)	15,73	2,84	9,17	22,29	5,53	0,0006
COBVEG(X2)	-3,99	1,33	-7,06	-0,92	-3	0,0171

El método de mínimos cuadrados aplicado al modelo:

$$PERDS_i = \beta_0 + \beta_1 \cdot INCL_i + \beta_2 \cdot COBVEG_i + \varepsilon_i$$

generó la siguiente ecuación estimada:

$$\hat{PERDS}_i = -2.67 + 15.73 \cdot INCL_i - 3.99 \cdot COBVEG_i$$

El valor estimado $\hat{\beta}_1 = 15.73$, significa que por cada aumento unitario en la inclinación, para cada valor particular de la cobertura, se espera un aumento promedio de la pérdida de suelo de $15.73 g/m^2$.
De manera análoga, el valor estimado $\hat{\beta}_2 = -3.99$, significa que por cada aumento unitario en la cobertura vegetal, para cada valor particular de la inclinación, se espera una disminución promedio de la pérdida de suelo de $3.99 g/m^2$.

Tanto la inclinación como la cobertura vegetal explican significativamente la pérdida de suelo (valores p = 0,0006 y 0,0171 respectivamente, para las pruebas de hipótesis respecto a los coeficientes de regresión parciales).

Al igual que en la regresión simple, se pueden calcular los valores predichos y los respectivos residuales. Por ejemplo, para una inclinación de 0.48 y una cobertura de 0.3 la pérdida de suelo predicha es $3.68 g/m^2$ y el residual respecto al valor observado es $0.10 g/m^2$

A continuación se presenta la tabla de análisis de varianza para este modelo:

Cuadro de Análisis de la Varianza (SC tipo I: secuencial)

Fuente de Variación	SC	gl	CM	F	valor p
Modelo de regresión	6,7	2	3,35	101,33	<0,0001
INCL(X1)	6,4	1	6,4	193,67	<0,0001
COBVEG(X2)	0,3	1	0,3	8,98	0,0171
Error	0,26	8	0,03		
Total	6,97	10			

Se observa que la fuente de variación "modelo", que corresponde al modelo en conjunto, explica una parte importante de la variación total ($6,7/6,97 = 0,96 = R^2$) y tiene asociados 2 grados de libertad (igual al número de variables regresoras). La prueba de hipótesis que se representa en la fila de la tabla denominada "Modelo"es:

Ho: $\beta_1 = \beta_2 = 0$ (ninguna de las variables explicativas consideradas en el modelo es significativa para explicar la variación en la respuesta)

H1: Al menos alguno de los parámetros es distinto de cero.

El cociente F de esta prueba es:

$F = CM_{REGRESIÒN} / CM_{ERROR} = 101,33$

El valor p asociado es < 0.05, en consecuencia se concluye que el modelo en conjunto explica significativamente la variación en la variable respuesta y al menos alguna de las variables regresoras es significativa. Es decir, ya sabemos que podemos predecir las pérdidas de suelo conociendo la inclinación de la superficie y/o la cobertura vegetal.

Las pruebas correspondientes a los siguientes renglones de la tabla nos permiten contestar primero si la inclinación es significativa para explicar la pérdida de suelo ($\beta_1 \neq 0$) y luego si la cobertura vegetal contribuye significativamente a la predicción de la pérdida de suelo una vez que la inclinación ya fue incluida en el modelo ($\beta_2 / x_1 \neq 0$, se lee beta dos, dado que X1 ya está en el modelo, es distinto de cero). Estas sumas de cuadrados se denominan secuenciales porque representan la porción de la variación total que es explicada al añadir al modelo cada nueva variable. Es decir que, en este ejemplo, incluir la cobertura vegetal en el modelo corresponde a una suma adicional de cuadrados (variación adicional explicada) de 0,3. Los valores de p asociados a ambas pruebas son menores a 0.05, en consecuencia se concluye con esa probabilidad de error que las dos variables explicativas son significativas. Las SC secuenciales de todas las variables consideradas totalizan la SC del modelo (en este caso SCX1 + SCX2= 6,4+0,3=6,7). Las pruebas de hipótesis asociadas a las sumas de cuadrados secuenciales se suelen denominar pruebas F secuenciales.

Cuando se calcula para cada variable explicativa la SC adicional que generaría al entrar última al modelo, se obtienen las denominadas SC parciales, las que se pueden ver para nuestro ejemplo en la tabla siguiente:

Cuadro de Análisis de la Varianza (SC tipo III)				parcial	
Fuente de Variación	SC	gl	CM	F	Valor p
Modelo de regresión	6,7	2	3,35	101,33	<0,0001
INCL(X1)	1,01	1	1,01	30,57	0,0006
COBVEG(X2)	0,3	1	0,3	8,98	0,0171
Error	0,26	8	0,03		
Total	6,97	10			

Se observa que la prueba F para el modelo en conjunto (primer renglón de la tabla) es idéntica a la de la tabla anterior, ya que se trata del mismo modelo. También la prueba F correspondiente a la variable X2 es igual en ambas tablas, ya que la SC secuencial de la última variable que entra al modelo es igual a la correspondiente SC parcial. Finalmente, se puede comprobar que las pruebas F parciales son idénticas a las pruebas t de significación de los parámetros que se presentaron en la tabla de los coeficientes: es fácil verificar que los valores del estadístico F son iguales a los cuadrados de los correspondientes valores del estadístico t y que los valores p asociados son idénticos. Esto es así porque $F_{1,\nu,1-\alpha} = (t_{\nu,1-\alpha/2})^2$. Es decir que las pruebas t de significación de los parámetros se refieren a su aporte cuando las demás variables regresoras ya están incluidas en el modelo. Se observa que las SC parciales, a diferencia de las secuenciales, no totalizan la SC del modelo completo al adicionarlas y es correcto que así sea. En este ejemplo, los valores F para las SC parciales indican que todas las variables consideradas son importantes en la ecuación de regresión; es decir que no es necesario considerar la eliminación de ninguna de ellas. Esto no siempre es así, como veremos a continuación.

7.2.9. El problema de la multicolinealidad

Es posible que en un conjunto de datos algunas variables regresoras resulten significativas en el procedimiento secuencial y no lo sea ninguna en el procedimiento parcial. Esto indica que la explicación de la variación que cada variable puede aportar ya ha sido explicada suficientemente por las demás y generalmente se debe a que existen correlaciones bastante altas entre las variables regresoras, problema denominado de ***multicolinea-***

lidad. Esta dificultad se pone de manifiesto en el ejemplo del estudio de pérdidas de suelo por erosión, al incluir la variable X3 en el modelo, que entonces queda expresado por la siguiente ecuación:

$$PERDS_i = \beta_0 + \beta_1 \cdot INCL_i + \beta_2 \cdot COBVEG_i + \beta_3 \cdot ABMAX_i + \varepsilon_i$$

A continuación se presenta la tabla de coeficientes de regresión parcial y sus estimaciones por intervalos de confianza.

Coeficientes de regresión y estadísticos asociados

Coeficiente	Estimador	EE	LI(95%)	LS(95%)	T	Valor p
constante	-0,19	1,81	-4,48	4,1	-0,1	0,9204
INCL(X1)	7,73	4,45	-2,79	18,26	1,74	0,1259
COBVEG(X2)	-2,39	1,34	-5,57	0,79	-1,78	0,1183
ABMAX(X3)	0,16	0,07	-0,02	0,33	2,12	0,0714

Se observa que todos los intervalos de confianza contienen al cero y en forma coherente con este resultado, ninguna de las hipótesis nulas que postulan que los parámetros son iguales a cero, pueden ser rechazadas (p= 0.12, 0.11 y 0.07 respectivamente). Es decir que los datos no aportan evidencias suficientes para afirmar que alguna de las variables propuestas explique parcialmente una porción significativa de la variación en pérdidas de suelo entre parcelas.

Sin embargo, en las correspondientes tablas de análisis de varianza (secuencial y parcial) para el modelo propuesto, que se presentan a continuación, se observa que hay evidencia suficiente para rechazar la H_0 del modelo en conjunto y de las dos primeras variables en forma secuencial. Esto coincide con los resultados del análisis del modelo con dos variables (X1 y X2):

Cuadro de Análisis de la Varianza (SC tipo I: secuencial)

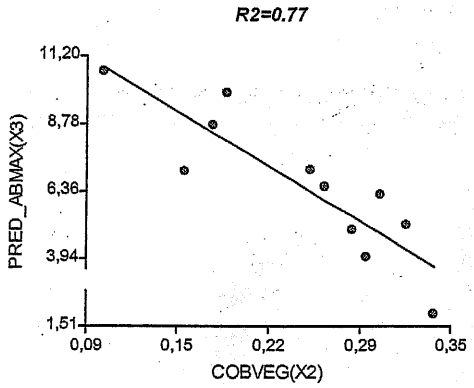
Fuente de Variación	SC	gl	CM	F	Valor p
Modelo de regresión	6,8	3	2,27	98,66	<0,0001
INCL(X1)	6,4	1	6,4	278,55	<0,0001
COBVEG(X2)	0,3	1	0,3	12,92	0,0088
ABMAX(X3)	0,1	1	0,1	4,51	0,0714
Error	0,16	7	0,02		
Total	6,97	10			

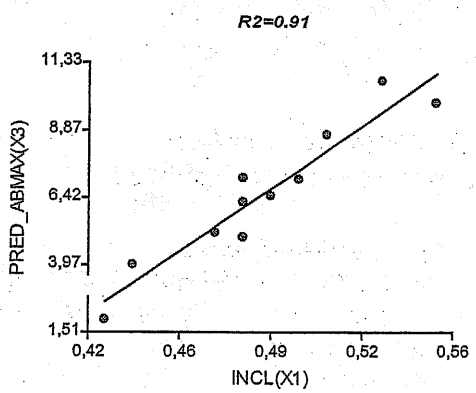
Cuadro de Análisis de la Varianza (SC tipo III: parcial)

Fuente de Variación	SC	gl	CM	F	Valor p
Modelo de regresión	6,8	3	2,27	98,66	<0,0001
INCL(X1)	0,07	1	0,07	3,02	0,1259
COBVEG(X2)	0,07	1	0,07	3,17	0,1183
ABMAX(X3)	0,1	1	0,1	4,51	0,0714
Error	0,16	7	0,02		
Total	6,97	10			

Se observa que al incorporar el diámetro de la mayor abertura de suelo desnudo (X3) como variable explicativa, los coeficientes de las otras dos variables (X1 y X2) ya no resultan significativamente distintos de cero (pruebas t de la primera tabla y equivalentes pruebas F de la última tabla). En otras palabras: ninguna de las variables explica por sí sola una porción importante de la variación cuando las otras dos ya fueron incluidas

en el modelo. Esto ocurre porque las tres variables explicativas están fuertemente asociadas entre sí, como se puede observar en los siguientes gráficos:





Los gráficos muestran que la variación del tamaño de la abertura máxima de suelo desnudo de las parcelas es explicada en un alto porcentaje tanto por la cobertura vegetal como por la inclinación de las parcelas. Por eso el modelo que incluye a las tres variables, si bien explica una proporción alta de la variación en las pérdidas de suelo (R^2= 6,8/6,97= 0.98) y en conjunto resulta significativo, no detecta a ninguno de los coeficientes parciales como significativo, ya que ninguna de las variables explica una porción interesante de la variación que no haya sido explicada por las otras dos. Intuitivamente percibimos que el modelo reducido (con 2 variables explicativas) es más adecuado que el modelo completo (con 3), pero además existen maneras formales de ponerlo a prueba.

Con este ejemplo se pretende sólo presentar una introducción al problema de la selección del "mejor" modelo, que es mucho más amplio y no será abordado en este curso. Algunos textos discuten este tema con mayor detalle y presentan varios procedimientos relacionados (Draper y Smith 1981, Kleinbaum, Kupper y Muller 1988).

7.2.10. Regresión polinómica

Un caso particular de regresión múltiple lo constituyen los modelos de mayor orden, donde una misma variable se incluye en el modelo elevada a distintas potencias. Por ejemplo, el siguiente es un modelo lineal de tercer orden con una variable explicativa:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i$$

El análisis de este modelo no presenta diferencias con el modelo general, ya que claramente la variable al cuadrado y la variable al cubo podrían ser consideradas distintas variables explicativas. Así, si denominamos a X_1^2 como X_2 y a X_1^3 como X_3, resulta un modelo idéntico al analizado previamente con tres variables explicativas. En consecuencia, todas las consideraciones expuestas para el modelo general de regresión múltiple se aplican de manera directa a la regresión polinómica.

Por ejemplo, en una planta procesadora de alimentos se estudia la aplicación de un conservante para jugos de frutas y se evalúa la concentración de una determinada sustancia en función del tiempo transcurrido desde la aplicación del conservante. Se preparan 18

soluciones y se mide la concentración de interés a las 6, 8, 10, 12, 14 y 16 horas transcurridas desde la aplicación. A continuación se presentan los datos obtenidos:

Solución	1	2	3	4	5	6	7	8	9
Tiempo (h)	6	6	6	8	8	8	10	10	10
Conc (mg/ml)	2.11	1.52	2.75	3.01	2.43	3.76	4.22	6.12	4.96

Solución	10	11	12	13	14	15	16	17	18
Tiempo (h)	12	12	12	14	14	14	16	16	16
Conc (mg/ml)	8.46	7.02	7.95	10.42	11.78	12.21	16.93	16.11	15.67

El modelo propuesto es:

$$CONC_i = \beta_0 + \beta_1 \cdot TIEMPO_i + \beta_2 \cdot TIEMPO_i^2 + \varepsilon_i$$

En la tabla siguiente se presentan los estimadores de los parámetros:

Coeficientes de regresión y estadísticos asociados

Coeficiente	Estimador	EE	LI(95%)	LS(95%)	T	valor p
const	4.68	1.87	0.69	8.66	2.5	0.0243
Tiempo (h)	-1.11	0.36	-1.88	-0.34	-3.07	0.0078
(Tiempo (h))^2	0.11	0.02	0.08	0.15	7.01	<0.0001

Todos los parámetros difieren significativamente de cero y la ecuación estimada es:

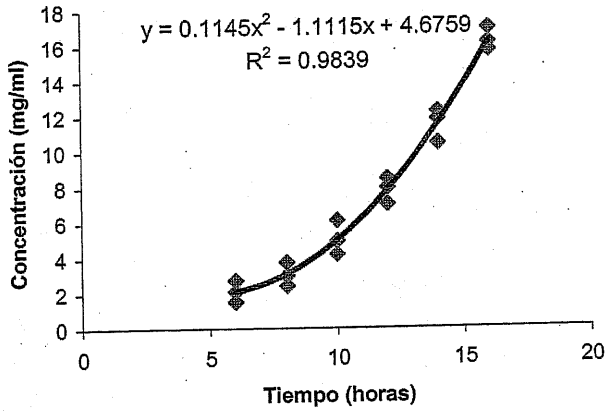
$$\hat{CONC}_i = 4.68 - 1.11 \cdot TIEMPO_i + 0.11 \cdot TIEMPO_i^2$$

La tabla de ANOVA asociada al modelo es:

Cuadro de Análisis de la Varianza (SC tipo I)

F.V.	SC	gl	CM	F	Valor p
Modelo	439.04	2	219.52	459.44	<0.0001
Tiempo (h)	415.56	1	415.56	869.75	<0.0001
(Tiempo (h))^2	23.48	1	23.48	49.14	<0.0001
Error	7.17	15	0.48		
Total	446.2	17			

Se rechaza H_0 para el modelo en conjunto y para cada uno de sus parámetros.
El coeficiente de determinación $R^2 = SC_R/SC_{TOTAL} = 0.98$, indica que el 98% de la variación de concentraciones entre soluciones es explicada por el tiempo transcurrido desde la reacción. El gráfico muestra los valores observados y los estimados por el modelo.



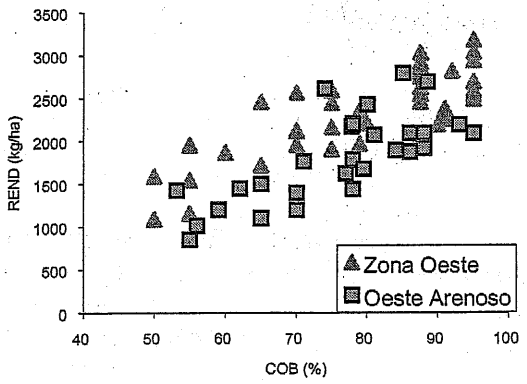
7.2.11. Regresión con variables explicativas categóricas

Las variables categóricas son aquellas que asumen valores arbitrarios (usualmente 0, 1, -1) que identifican distintas categorías de una variable nominal, como distintos grupos, tratamientos o, de manera más general, situaciones experimentales. Es decir que estas

variables explicativas no describen niveles de medida sino que son fijadas arbitrariamente por el investigador para identificar grupos de observaciones. En el contexto del análisis de regresión su aplicación más importante se relaciona con la posibilidad de comparar los modelos de regresión ajustados a esos grupos de observaciones.

Retomando el ejemplo de rendimientos de girasol del comienzo del capítulo, recordemos que se había registrado tanto la cobertura del cultivo al momento de floración como los rendimientos, en lotes cultivados en una zona que ahora denominaremos Zona Oeste. La experiencia también se realizó simultáneamente en otro ambiente de suelos más arenosos y menores precipitaciones que denominaremos Oeste Arenoso.

El gráfico muestra con símbolos diferentes los datos obtenidos en ambas zonas. El modelo propuesto para la Zona Oeste se puede ampliar para abarcar las dos zonas mediante un modelo de regresión múltiple con variables categóricas, como se verá a continuación.



Se genera una variable Z a la que se le asignan dos valores posibles:

Z_i= 0 para los lotes de la Zona Oeste
Z_i= 1 para los lotes de Oeste Arenoso

El modelo que describe la relación entre cobertura en floración y rendimiento para ambas zonas es el siguiente:

$$REND_i = \beta_0 + \beta_1 \cdot COB_i + \beta_2 \cdot Z_i + \beta_3 \cdot COB_i Z_i + \varepsilon_i$$

Este modelo resulta en dos ecuaciones de modelos de regresión simple, una para cada zona, cuando Z_i asume los valores 0 y 1 que identifican a los lotes de cada ambiente:

$Z_i = 0$ ➔ Zona Oeste $\qquad REND_i = \beta_0 + \beta_1 \cdot COB_i + \varepsilon_i$

$Z_i = 1$ ➔ Oeste Arenoso $\qquad REND_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot COB_i + \varepsilon_i$

Se observa que la pendiente para la Zona Oeste está representada por el parámetro β_1 y para el Oeste Arenoso, por la suma de parámetros $\beta_1 + \beta_3$. De esta forma, la prueba de hipótesis para el parámetro β_3 equivale a probar si ambas rectas tienen la misma pendiente, por eso se la denomina ***prueba de paralelismo***. Sólo si no se rechaza la hipótesis de paralelismo, interesa investigar si ambas rectas tienen también igual ordenada al origen ($\beta_2 = 0$), que es la denominada ***prueba de coincidencia***. De no rechazar esta hipótesis se concluye que no hay evidencias para ajustar modelos distintos a ambos grupos de casos.

En el ejemplo de los lotes de girasol el modelo estimado resulta:

$$\hat{RÊND}_i = 49.83 + 28.44 \cdot COB_i - 924.11 \cdot Z_i + 6.96 \cdot COB_i Z_i$$

Este modelo resulta en dos ecuaciones de modelos de regresión simple, una para cada zona, cuando Z_i asume los valores 0 y 1 que identifican a los lotes de cada ambiente:

$Z_i = 0$ ➔ Zona Oeste $\quad \hat{RÊND}_i = 49.83 + 28.44 COB_i$

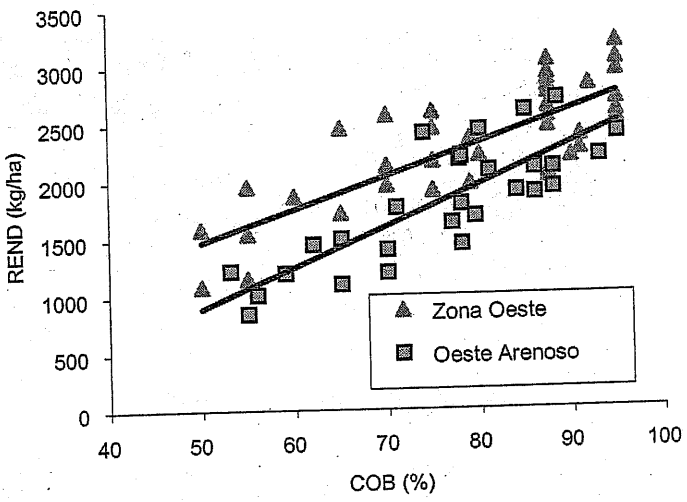
$Z_i = 1$ ➔ Oeste Arenoso $\quad \hat{RÊND}_i = -874.28 + 35.4 \cdot COB_i$

En la tabla siguiente se presentan los estimadores de los parámetros:

Coeficiente	Estimador	EE	LI(95%)	LS(95%)	T	valor p
b0	49.83	280.67	-510.39	610.06	0.18	0.8596
b1	28.44	3.49	21.48	35.41	8.15	<0.0001
b2	-924.11	481.51	-1885.2	36.97	-1.92	0.0592
b3	6.96	6.18	-5.37	19.29	1.13	0.2642

La prueba de hipótesis de paralelismo no se rechaza (p=0.2642), ni tampoco se rechaza la prueba de coincidencia entre los modelos (p=0.0592).
Se concluye que el rendimiento responde de forma similar al incremento de la cobertura en ambas zonas (las pendientes de las respuestas no difieren significativamente), y que los rendimientos esperados de la Zona Oeste son similares a los del Oeste Arenoso. Por lo tanto, no existen evidencias suficientes para ajustar modelos distintos a ambos grupos de casos. Notar que las ordenadas al origen no se interpretan como rendimiento esperado para el valor cero de cobertura en floración (lo cual sería bastante absurdo) porque el rango de cobertura estudiado no abarca el valor cero: comienza en coberturas superiores a 40%.

El siguiente gráfico muestra las funciones ajustadas para cada zona y las observaciones. El lector ya habrá notado que el modelo estimado para la Zona Oeste coincide con el modelo de regresión simple ajustado para los mismos datos al comienzo del capítulo. La misma coincidencia se encuentra si se ajusta un modelo de regresión simple para los lotes del Oeste Arenoso. La ventaja que ofrece la regresión múltiple con variables categóricas es la posibilidad de comparar los parámetros de ambas regresiones.



A continuación se presenta la tabla de ANOVA:

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	15195634.3	3	5065211.4	52.62	<0.0001
B1	6399032.97	1	6399033	66.47	<0.0001
B2	386305.26	1	386305.26	4.0129	0.0592
B3	122066.83	1	122066.83	1.27	0.2642
Error	6449812.52	67	96265.86		
Total	21645446.9	70			

$R^2 = SC_R/SC_{TOTAL} = 15195634.3/\ 21645446.9 = 0.702$

Es decir que el modelo explica el 70% de la variación en los rendimientos de los lotes. La prueba de hipótesis para el modelo en conjunto ($p<0.0001$) indica que alguno de los parámetros es significativamente distinto de cero. Las pruebas individuales coinciden con las pruebas t de la tabla de coeficientes y estimadores, como fuera explicado anteriormente.

7.2.12 El análisis de covarianza y la regresión con variables categóricas

El análisis de covarianza (ANCOVA), al igual que el análisis de la varianza, se utiliza para poner a prueba los efectos de dos o más tratamientos sobre una variable respuesta, pero su particularidad radica en que permite controlar la posible influencia sobre la respuesta de otra variable cuantitativa observada, con el objetivo de reducir el error experimental. En consecuencia, esta es una técnica que ocasionalmente resulta ventajosa para mejorar la potencia de los experimentos. En el siguiente ejemplo se describe una de las situaciones en que el análisis de covarianza podría resultar de gran utilidad.

En un estudio de comparación de raciones, donde la variable respuesta (Y) es la ganancia de peso, se sospecha que las diferencias en peso inicial podrían generar una variación no debida a los tratamientos en la variable respuesta. El investigador no puede variar a voluntad el peso inicial de los animales (X) pero sí puede registrarlo. Tampoco está interesado en estudiar el efecto del peso inicial sobre la ganancia de peso, pero no desea que éste actúe oscureciendo el efecto diferencial de las raciones. La variable X se denomina *covariable*. El análisis de covarianza consiste en ajustar la variable respuesta observada por el efecto de la covariable, antes de comparar las medias de los tratamientos. Si no se realizara el ajuste, el efecto no registrado de la covariable sobre la variable respuesta podría aumentar el cuadrado medio del error y dificultar la detección de diferencias verdaderas entre las medias.

Conceptualmente, utilizar una covariable en el modelo es equivalente al uso de bloques para reducir el error experimental y consecuentemente aumentar la potencia del ensayo. Particularmente cuando la variable de bloqueo es cuantitativa, la delimitación de los bloques puede resultar muy subjetiva. En cambio el ANCOVA no requiere esa clasificación en grupos discretos, ya que permite remover la variación en Y que se encuentra asociada a X utilizando las medidas conjuntas de ambas variables cuantitativas en cada unidad experimental. En el estudio de raciones, por ejemplo, la heterogeneidad en ta-

maño inicial de los animales podría ser controlada en un diseño en bloques repartiendo los animales en grupos de tamaño (un bloque de animales pequeños, otro de medianos y otro de grandes). En cambio en el ANCOVA, el peso inicial de cada animal intervendría directamente en el análisis sin requerir una clasificación previa en grupos de tamaño.

El procedimiento del ANCOVA es similar al Análisis de Regresión con Variables Categóricas (ARVC). Esto es así porque ANCOVA y ARVC tratan con modelos generalizados, es decir modelos que combinan en su porción explicativa tanto variables cuantitativas como nominales. Efectivamente, el procedimiento de ARVC presentado en el capítulo anterior sirve perfectamente para analizar datos de ANCOVA, como se verá más adelante. La principal distinción que justifica su presentación como modelos diferentes, radica en los **objetivos** que guían cada estudio:

- ✓ En el ANCOVA la pregunta principal está asociada con las medias de los tratamientos (variables nominales) y la relación de la respuesta con otra variable cuantitativa independiente de los tratamientos que puede ser aprovechada para aumentar la potencia de las comparaciones.

- ✓ En el ARVC la pregunta principal se refiere a la relación funcional de la respuesta con una variable cuantitativa y además interesa conocer si esa relación difiere entre los tratamientos (identificados mediante variables categóricas).

Un ejemplo del primer caso sería un trabajo que pretende comparar los **costos mensuales** de mantenimiento de tractores (variable respuesta) utilizando distintas **calidades de aceite** (tratamientos), sin dejar de considerar como fuente de variabilidad adicional los **kilómetros recorridos** por cada tractor (covariable). Un ejemplo del segundo caso podría ser el estudio de la variación en el **volumen de madera** (variable respuesta) en función del **diámetro** de troncos de *Pinus elliotti* (variable explicativa cuantitativa) y cuando además se desea indagar si la relación en cuestión difiere para distintas **variedades** de *Pinus elliotti* (variable categórica).

El modelo de ANCOVA para un diseño completamente aleatorio es:

$$y_{ik} = \mu + \tau_i + \beta_1(x_{ik} - \bar{x}_{..}) + \varepsilon_{ik}$$

Donde el valor de la respuesta cuando se aplicó el tratamiento *i* a la unidad experimental *k*, se considera generado por:
μ, que es la media general.
τ_i, que es el efecto del tratamiento i,
$\beta_1(x_{ik} - \bar{x})$, que es el efecto de la relación entre *Y* y *X* producido por la desviación del valor de la covariable respecto a la media.
ε_{ik}, que son errores independientes con distribución $N(0; \sigma^2)$.

$i = 1, ..., t$ $\qquad$ $k = 1, ..., r$

Además de los supuestos usuales de independencia, homogeneidad de varianzas y distribución normal de los errores, el modelo de ANCOVA tiene como supuesto adicional que el coeficiente de regresión: β_1, sea el mismo para los distintos tratamientos. Es

decir que la pendiente de la relación entre la variable respuesta y la covariable no debe diferir entre tratamientos. En términos del ejemplo del comienzo del capítulo, implica suponer que la influencia del peso inicial sobre el aumento de peso no difiere entre animales que recibieron distintas raciones. De hecho, si la relación entre la covariable y la respuesta resultara diferente entre los tratamientos, esa diferencia entre las relaciones pasaría a ser el resultado más relevante del estudio y perdería todo interés la comparación entre las medias de los tratamientos (resultando el ARVC el contexto más adecuado para el análisis de los datos).
El mismo modelo escrito en términos de ARVC para t = 2, queda expresado como:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 Z_i + \varepsilon_i$$

Donde Z_i es la variable categórica que toma valor 0 (cero) para un tratamiento y 1 para el otro.

Cuando $Z_i = 0$ (Tratamiento 1) el modelo resulta: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Cuando $Z_i = 1$ (Tratamiento 2) el modelo resulta:

$$y_i = \beta_0 + \beta_2 + \beta_1 x_i + \varepsilon_i$$

El lector avezado ya habrá notado que la única diferencia entre este modelo y el presentado en el capítulo anterior es la ausencia del parámetro β_3, el mismo que permitía poner a prueba la diferencia de pendientes entre tratamientos. Ese parámetro no aparece en este modelo que asume pendientes iguales.

La equivalencia entre los parámetros de los modelos de ANCOVA y ARVC es (sólo para curiosos):

ANCOVA	ARVC
$\mu + \tau_1 - \beta\overline{x_{..}}$	β_0
β_1	β_1
$\tau_2 - \tau_1$	β_2

7.2.13 Las hipótesis EN ANCOVA Y EN ARVC

Las preguntas de interés son:

1) ¿Existen diferencias en la respuesta media entre los tratamientos?
2) ¿Influye la covariable sobre la variable respuesta?

La hipótesis nula asociada a la primera pregunta es:

En ANCOVA: $H_o : \tau_i = 0$

En ARVC: $H_0 : \beta_2 = 0$

La hipótesis nula asociada a la segunda pregunta es para ambos modelos:

$$H_0 : \beta_1 = 0$$

Cada vez que no se pueda asegurar con información previa el supuesto de igual pendiente para los distintos tratamientos, se incluye el parámetro β_3 en el modelo para ponerlo a prueba (en un ARVC, ver Ejemplo 2), en este sentido el ARVC es más versátil. Sin embargo, cuando los diseños son un poco complicados (modelos mixtos o factoriales de gran dimensión) el ANCOVA ofrece un contexto más apropiado para el análisis. Conceptualmente el ANCOVA es muy semejante a realizar un análisis de regresión de la variable respuesta sobre la covariable y luego comparar las medias de los residuales (la variación en la respuesta no explicada por la covariable) entre los distintos tratamientos con un ANOVA. Sin embargo, no se recomienda reemplazar el análisis de covarianza por este procedimiento (García Berthou 2001) ya que no produce resultados idénticos, entre otras cosas porque el ANOVA de los residuales no toma en cuenta el grado de libertad utilizado para estimar el coeficiente de la regresión.

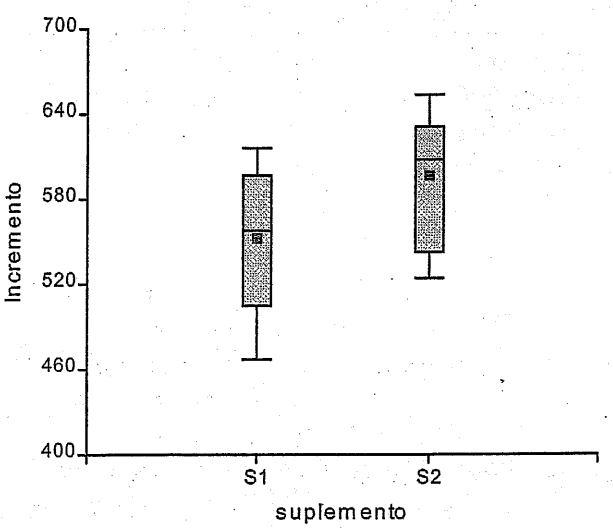
EJEMPLO 1

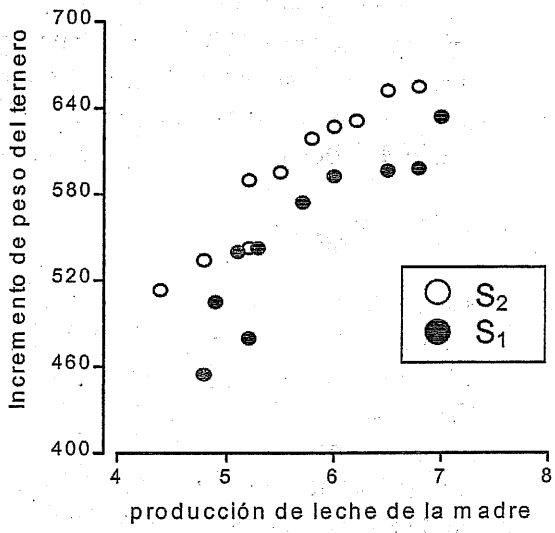
La suplementación diferencial de terneros al pie de la madre, práctica denominada Creep-feeding, tiene como objetivo aumentar la tasa de crecimiento de los terneros lactantes en el último período antes del destete. En un experimento de comparación entre dos raciones suplementarias para ser utilizadas en programas de alimentación de terneros al pie de la madre, se registró el aumento diario de peso de 20 terneros alimentados con cada uno de dos suplementos que denominaremos S_1 y S_2. Debido a que el aumento de peso alcanzado por cada ternero lactante sería afectado no sólo por la ración sino también por la producción de leche de cada madre (que varía con su capacidad genética, con la disponibilidad de pasturas y con su historia nutricional previa), se registró también la producción diaria de leche de la madre para ser utilizada como covariable en el análisis.

Suplemento S_1		Suplemento S_2	
ternero	madre	ternero	madre
Aumento de peso	Producción de leche	Aumento de peso	Producción de leche
(g/día)	(litros/día)	(g/día)	(litros/día)
634	7	655	6.8
598	6.8	653	6.5
597	6.5	632	6.2
593	6	628	6
574	5.7	619	5.8
542	5.3	596	5.5
540	5.1	590	5.2
505	4.9	542	5.2
480	5.2	534	4.8
455	4.8	513	4.4

En los siguientes gráficos se representa en las ordenadas el incremento diario de peso de los terneros en los dos tratamientos de suplementación. En el box-plot de la izquierda se observa que la gran variabilidad en las respuestas dentro de los tratamientos ensombrece a las posibles diferencias entre ellos. En el gráfico de dispersión de la derecha se observa que gran parte de esa variabilidad es explicada por la variación en la producción de leche de las madres y que para cada nivel de esta covariable, uno de los tratamientos

supera al otro en casi todos los casos observados. A partir de estas observaciones parece esperable que al remover la influencia de la covariable sobre la respuesta, resulte más fácil detectar diferencias en el efecto de los tratamientos de suplementación, si las hubiere.





Efectivamente, como se observa en la siguiente tabla de Análisis de la Varianza, el experimento aporta suficiente evidencia (P=0.0001) en contra de la hipótesis que postula que los suplementos no presentan efecto diferencial sobre el aumento de peso de los terneros ($H_o : \tau_i = 0$).

Tabla de Análisis de la Varianza (ANCOVA)

F.V.	SC	gl	CM	F	valor P	Coef
Suplemento	12554.54	1	12554.54	28.62	0.0001	
Prod leche	45882.58	1	45882.58	104.61	<0.0001	64.45
Error	7456.62	17	438.62			
Total	63196.00	19				

En la misma tabla se observa que la relación lineal entre la respuesta y la covariable es significativa (se rechaza $H_0 : \beta_1 = 0$, P<0.0001) y la pendiente estimada de esa relación es 64.45 g de incremento diario de peso del ternero por litro de producción de leche de la madre.

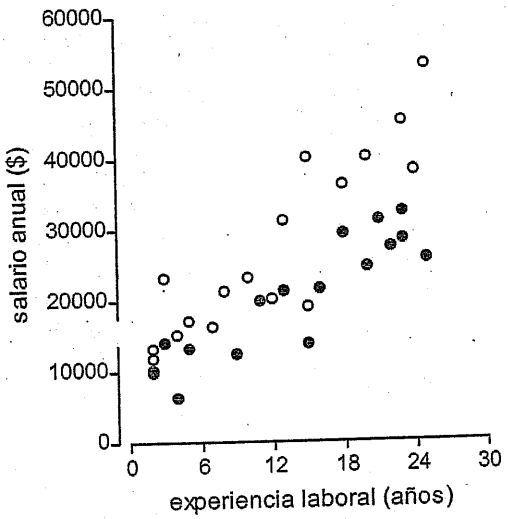
Queda como ejercicio para el lector comprobar que: 1) en ARVC bajo el supuesto de iguales pendientes se obtienen exactamente los mismos resultados, 2) en ARVC con el modelo completo no se rechaza $H_0 : \beta_3 = 0$ y 3) si se realiza un ANOVA sin considerar el efecto de la covariable, los datos no aportan evidencias para concluir que las medias de aumento de peso diario de los terneros difieren significativamente entre los tratamientos de suplementación (P= 0.085).

EJEMPLO 2

En un estudio de comparación entre los salarios que perciben los profesionales varones y las profesionales mujeres de una ciudad, se selecciona una muestra al azar de 17 profesionales de cada género y se registra su salario anual en pesos (variable *Y*). El investigador procura averiguar si los salarios que perciben los varones son en promedio superiores a los que perciben las mujeres, pero previendo que el efecto de la experiencia laboral pudiera confundirse con el efecto del género sobre el salario, también registra la cantidad de años que cada individuo lleva ejerciendo la profesión (variable *X*). En principio no existe un interés particular en estudiar el efecto de la experiencia laboral sobre los salarios, sólo se desea evitar que este efecto quede mezclado con el efecto del género. A diferencia del ejemplo anterior, no existe información previa que permita sostener el supuesto de igual pendiente entre tratamientos para la relación funcional entre la variable respuesta y la covariable. En consecuencia se plantea el modelo completo en ARVC.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 Z_i + \beta_3 Z_i x_i + \varepsilon_i$$

En el gráfico se representan los salarios de los varones con símbolos vacíos y los de las mujeres con símbolos llenos. Se observa que a mayor experiencia laboral los salarios son mayores para ambos géneros.



Antes de indagar si los salarios de los varones superan en promedio a los de las mujeres, corrigiendo a ambos por la experiencia laboral, se debe someter a prueba si la pendiente de la relación funcional entre experiencia y salario no difiere significativamente entre géneros ($H_0 : \beta_3 = 0$).

Si se rechazara esta hipótesis no sería válida la corrección del ANCOVA, ya que este modelo asume la misma pendiente para los distintos tratamientos.

En la tabla siguiente se presentan los estimadores de los parámetros:

Coeficientes de regresión y estadísticos asociados

Coef.	Estimador	EE	LI(95%)	LS(95%)	T	valor P
b0	9637.06	2147.87	5250.52	14023.59	4.49	0.0001
b1	1445.97	149.64	1140.37	1751.57	9.66	<0.0001
b2	-2024.13	3155.97	-8469.49	4421.22	-0.64	0.5262
b3	-553.44	213.45	-989.38	-117.51	-2.59	0.0146

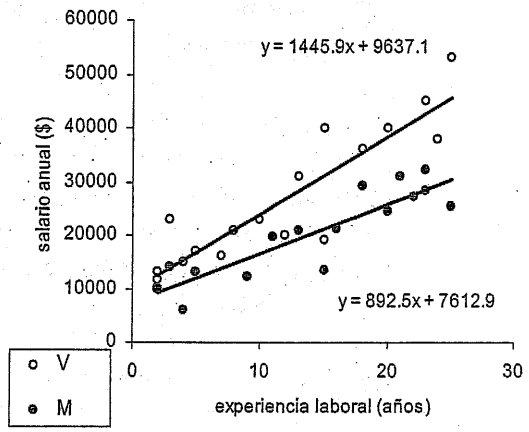
El modelo estimado es:

$$\hat{SAL}_i = 9637.06 + 1445.97 \cdot EXPE_i - 2024.13 \cdot Z_i - 553.44 \cdot EXPE_i \cdot Z_i$$

Este modelo resulta en dos ecuaciones de modelos de regresión simple, una para cada género, cuando Z_i asume los valores 0 y 1:

$Z_i = 0$ ➔ Varones $SAL_i = 9637.06 + 1445.97 \cdot EXPE_i$

$Z_i = 1$ ➔ Mujeres $\hat{SAL}_i = 7612.93 + 892.53 \cdot EXPE_i$



Se rechaza la hipótesis de igualdad de pendientes (P=0.0146). También se rechaza la hipótesis que postula que la experiencia laboral no explica la variabilidad en los salarios ($\beta_1 = 0$, P<0.0001). En cambio, los datos no aportan evidencias para afirmar que los salarios medios difieren entre géneros.

A continuación se presenta la tabla de ANOVA:

Cuadro de Análisis de la Varianza (SC tipo III)

F.V.	SC	gl	CM	F	p-valor
Modelo	3417211477.00	3	1139070492.33	50.55	<0.0001
b1	2104070809.96	1	2104070809.96	93.38	<0.0001
b2	9268971.64	1	9268971.64	0.41	0.5262
b3	151479665.17	1	151479665.17	6.72	0.0146
Error	675990875.94	30	22533029.20		
Total	4093202352.94	33			

$R^2 = SC_R/SC_{TOTAL} = 3417211477/\ 4093202352 = 0.83$

Es decir que el modelo explica el 83% de la variación en los salarios. La prueba de hipótesis para el modelo en conjunto (p<0.0001) indica que alguno de los parámetros es significativamente distinto de cero. Las pruebas individuales coinciden con las pruebas t de la tabla de coeficientes y estimadores. Se concluye que en la ciudad estudiada la principal diferencia entre los salarios que perciben los profesionales de ambos géneros radica en que los salarios de los varones aumentan más en promedio que los de las mujeres por cada año de aumento en su experiencia laboral.

7.2.14 Independencia de la covariable respecto a los tratamientos

Idealmente la covariable debería ser medida antes de que los tratamientos sean aplicados a las unidades experimentales, o al menos, se debería garantizar que no sea afectada por los tratamientos. En el ejemplo 1 la estrategia que garantiza que los tratamientos (raciones) no afecten a la covariable (producción de leche de las madres) es la forma de suministrar el alimento: en comederos ubicados en corrales a los que solo tienen acceso

los terneros. En algunos experimentos esta condición no se puede cumplir y, en consecuencia, los resultados del análisis se deben interpretar con precaución para no arribar a conclusiones extremadamente ingenuas. Por ejemplo, supongamos que en un experimento que evalúa el impacto de la presencia de distintas especies de malezas sobre el rendimiento de un cultivo de verano, se registra como covariable la *densidad final* lograda por el cultivo (número de plantas/m^2). Si resultara que no se encuentran diferencias significativas entre los rendimientos corregidos por la densidad, se podría concluir que las malezas no afectan diferencialmente al cultivo. Sin embargo, es posible que la reducción del rendimiento ocasionada por la presencia de las malezas se deba justamente a fallas en la emergencia o en la supervivencia de las plantas del cultivo, cuya primera evidencia se manifiesta en variaciones de la densidad. En este caso, al corregir los rendimientos por la densidad, en lugar de reducir el error experimental se estaría reduciendo la variación de la respuesta debida a los tratamientos, disminuyendo así la capacidad del experimento de detectar diferencias verdaderas.

7.2.15. Los residuos como elemento de diagnóstico

En los modelos de regresión, igual que en los otros modelos estadísticos que abarca este curso, los residuales son la porción de la variación no explicada por el modelo:

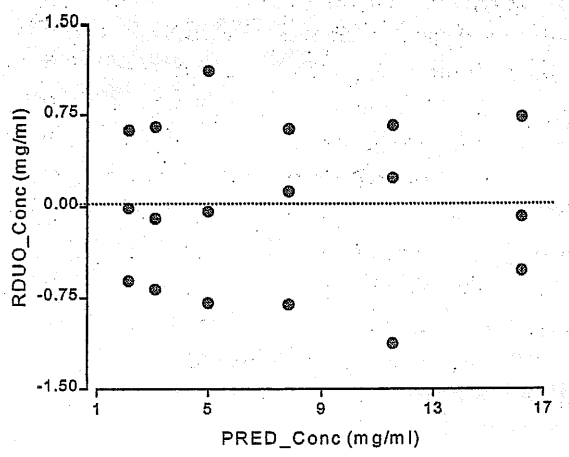
$$e_i = Y_i - \hat{Y}_i$$

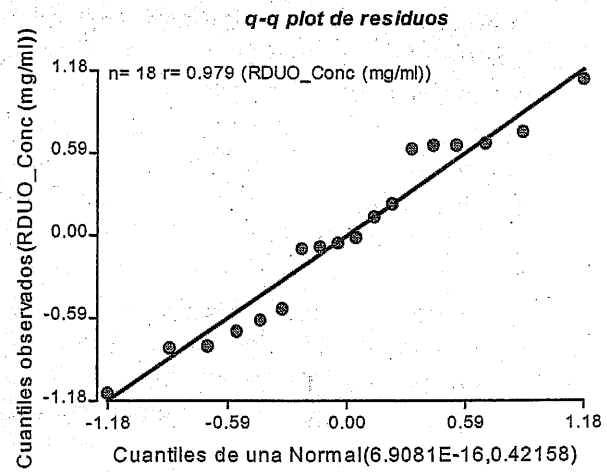
como tales, constituyen la única componente aleatoria de la variable respuesta. En consecuencia, el estudio de la distribución de los residuales nos permite contestar preguntas acerca de la distribución de la variable respuesta. Al respecto, las cuestiones principales a indagar se relacionan con el cumplimiento de los supuestos del modelo para que las pruebas de hipótesis asociadas tengan validez. En el caso particular del análisis de regresión, los residuales también pueden mostrar tendencias con alguna variable no incluida en el modelo o respuestas no lineales que no fueron consideradas y podrían mejorar el ajuste.

Recordemos que el modelo de regresión tiene como supuestos que las observaciones de la variable respuesta Y_i provienen de distribuciones de probabilidad cuyas medias son $E(Y_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_{p-1} X_{p-1\,i}$ y cuyas varianzas σ^2, son la misma para todos los niveles de X. Además cualesquiera dos observaciones Y_i y Y_j no están correlacionadas entre sí. Estos supuestos son necesarios para la estimación de parámetros. Además, para que las pruebas de hipótesis sean válidas se agrega el supuesto de distribución normal.

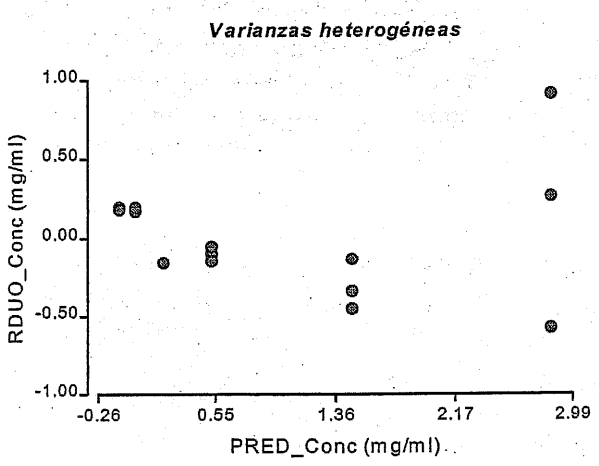
El supuesto de independencia entre las observaciones se debe garantizar en la etapa de colección de los datos (muestreo o diseño). En el ejemplo de regresión polinómica que vimos anteriormente, si en lugar de preparar 18 soluciones independientes y asignar al azar tres de ellas para cada hora de registro de la concentración, se hubieran preparado sólo tres soluciones y los registros a las distintas horas se hubieran realizado sobre las mismas tres soluciones, las observaciones correspondientes a la misma solución en distintos momentos estarían correlacionadas. Lo mismo en el ejemplo de los lotes de girasol en las distintas zonas, si por comodidad se toman conjuntos de varios lotes ubicados en los mismos establecimientos, estos no aportan observaciones independientes.

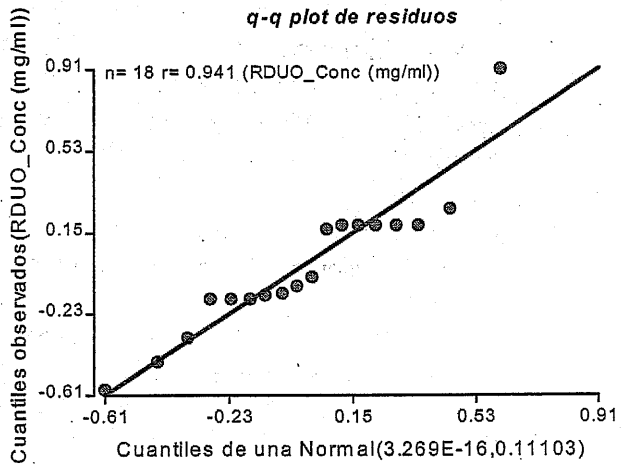
Los supuestos de homogeneidad de varianzas y normalidad se exploran mediante gráficos de residuales como los siguientes:





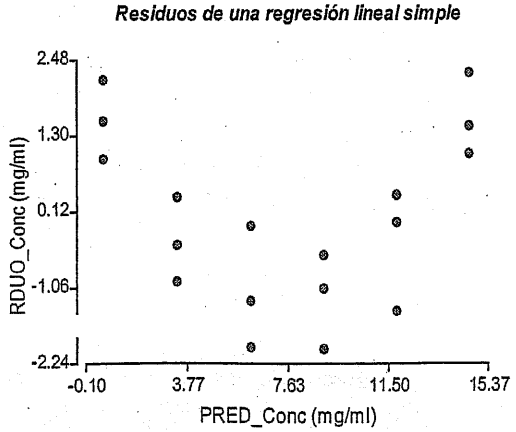
Los gráficos anteriores corresponden al mismo ejemplo de las concentraciones, en el cual los residuales se comportan de acuerdo a lo esperado. Si, en cambio, se obtuviera unos gráficos como los siguientes, habría que pensar seriamente en transformar la variable respuesta para ver si en otra escala presenta homogeneidad de varianzas y distribución más cercana a la normal.



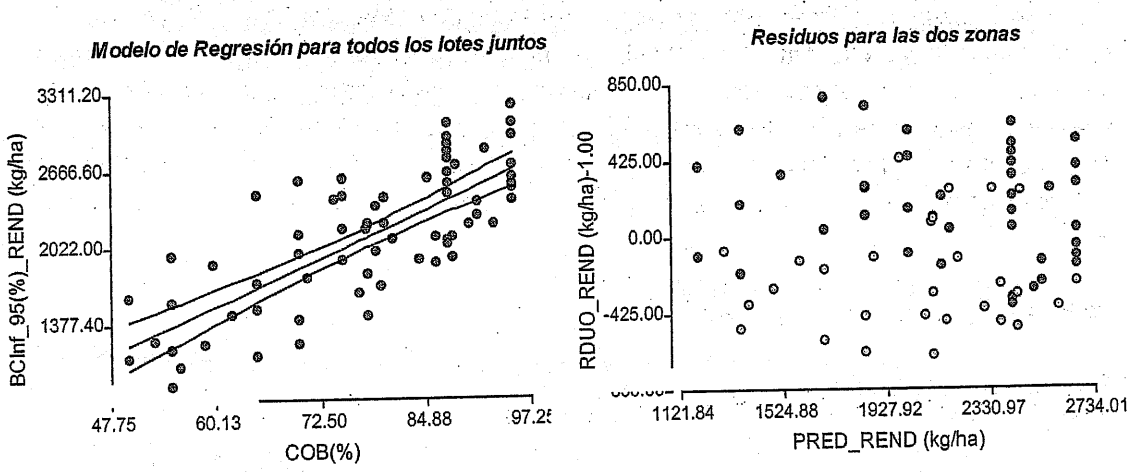


En relación con la adecuación del modelo, los residuales pueden dar señales interesantes como se verá en los siguientes casos.

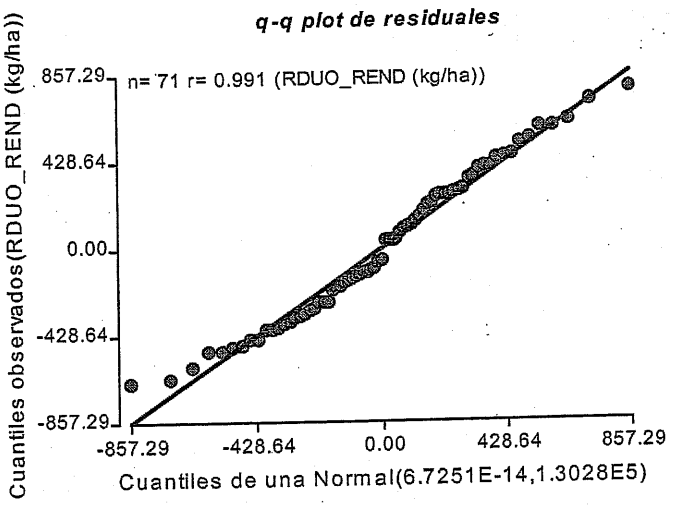
En el mismo ejemplo de las concentraciones de las soluciones, si se ajusta un modelo de regresión simple (sin incluir en el modelo la variable tiempo elevada al cuadrado), se obtiene este gráfico de residuales, donde se pone de manifiesto que resta explicar una parte de la variación de las concentraciones que presenta tendencia en forma de parábola (función cuadrática).



De manera análoga, en el problema de predicción de rendimientos de cultivos de girasol mediante los valores de cobertura en el momento de floración, si se ajusta un modelo para todos los lotes en conjunto, sin diferenciar por zonas, se obtiene la ecuación estimada: $\hat{Y} = -433.35 + 32.57\ X_i$, cuya representación gráfica y residuales se pueden observar en las siguientes figuras:



El gráfico anterior de residuales, donde se identifica con distintos colores la zona a la que pertenece cada uno de los lotes, pone de manifiesto que el modelo conjunto subestima los rendimientos para una de las zonas (los puntos oscuros están casi todos en la parte superior del gráfico) mientras los sobreestima para la otra (los puntos claros se encuentran mayormente por debajo del cero). En este caso los residuales sirven como advertencia para considerar la introducción en el modelo de una variable categórica que diferencie a los modelos de regresión para las distintas zonas.



Al mismo tiempo aquel gráfico de residuales muestra que no hay problemas serios de heterogeneidad de varianzas ya que para los distintos valores predichos la amplitud de los residuos es semejante. En el q-q plot que se presenta a la izquierda, que fue confeccionado con los mismos residuos, también se observa que no hay problemas respecto al supuesto de normalidad.

7.2.16 Comentarios adicionales

El análisis de regresión múltiple es una herramienta poderosa para estudiar las relaciones entre variables, pero debe ser utilizada con precaución. La importancia relativa de las variables candidatas a ser explicativas de una determinada respuesta debe fundamentarse antes en un marco de conocimiento teórico acerca del sistema bajo estudio que en pruebas de ensayo y error sobre grandes conjuntos de datos, donde pueden aparecer relaciones estadísticamente "significativas" pero imposibles de interpretar o carentes de sentido. Asimismo se debe tener en cuenta que para las pruebas de hipótesis el modelo propuesto es un supuesto. En otras palabras, no se puede "buscar el mejor modelo" mediante métodos de selección de variables y poner a prueba hipótesis acerca del modelo con los mismos datos, sino que se requieren conjuntos independientes de observaciones para realizar cada uno de esos pasos.

Bibliografía recomendada

Draper, N.R. & H. Smith (1981). Applied Regression Analysis. 2nd edition. Wiley, New York.

Kleinbaum, D.G., L.L. Kupper & K. E. Muller (1988). Applied Regression Analysis and other Multivariate Methods. Duxbury Press, Belmont, California.

Neter, J., W. Wasserman & M.H. Kutner (1985). Applied Linear Regression Models. Irwin, Homewood, Illinois.

Kuehl, R. 2001. Diseño de experimentos. Segunda edición. Thomson. Capítulo 17.

Montgomery, D. C. 1991. Diseño y análisis de experimentos. Grupo Editorial Iberoamérica. Capítulo 17.

Steel, R. G. y J. H. Torrie. 1985. Bioestadística: Principios y Procedimientos. Segunda Edición. McGraw-Hill. Capítulo 17.

Underwood, A. J. 1997. Experiments in Ecology. Cambridge University Press, Cambridge. Capítulo 13.